Author’s response to reviews

Title: Comparative efficacy of placebos in short-term antidepressant trials for major depression: A secondary meta-analysis of placebo-controlled trials

Authors:

Lisa Holper (lisa.holper@bli.uzh.ch)
Michael P. Hengartner (heng@zhaw.ch)

Version: 2 Date: 21 Aug 2020

Author’s response to reviews:

Dear Dr. Bousman,

Thank you very much for sending us the comments of the reviewers and for giving us the possibility to resubmit. All reviewer comments have been addressed and corresponding changes have been made throughout the manuscript. Please let us know if there are further questions.

Looking forward to hearing from you,
Sincerely
Lisa Holper

Reviewer 1:
This paper is a timely and very interesting manuscript. It describes an indirect comparison of placebos used in RCTs of major depressive disorders. Its results are counter-intuitive and challenge the validity of such large networks of evidence. This is of course though provoking and all the more important providing results are not over interpreted.
I must commend the authors for doing this research, for re-using smartly a very important database gathered by Cipriani and his colleagues. I must commend Cipriani and colleagues for sharing their dataset too.
Authors identify differences between certain placebos (amitriptyline placebo and trazodone placebo) and conclude that these differences are due to unblinding that should be more marked in trazodone and amitriptyline studies because the drugs caries more frequent adverse events.
Answer: The authors thank the reviewer for the positive response. All reviewer comments have been addressed and corresponding changes have been made throughout the manuscript

I have no comment on the statistical method used in the paper. I would only suggest authors to share their code publicly for reproducibility purpose.
Answer: The code of the unadjusted and covariate adjusted Bayesian models have been added to the supplement.
I have however very important conceptual/methodological comments that need to be addressed. These comments do not call for changes in the methods nor in the analysis but call for adding some uncertainty and more balance in the discussion.

While the results observed are supportive of the hypotheses of the authors (i.e. unblinding), there are many alternative explanations, even if authors tried to address as best as possible these limitations:

- First it must be clear that the hypotheses of authors were set a priori: was there a protocol (YES/NO)? was there a pre-registration (YES/NO, e.g. on the open science framework)? Authors must make these two points clear in the manuscript. If the answer is no to both points, it must be written in the methods section and the authors should state this as a limitation of their study in the discussion. If the answer is yes, the exact chronology between protocol registration and analyses should be transparently described.

**Answer:** The analysis was not based on a written protocol and there was no pre-registration. This information has been added to the manuscript at the beginning of the methods section (page 6): “The exploratory analysis was not based on a written protocol, but followed the findings of Naudet and colleagues (18).” And the beginning of the limitation section (page 14): “A limitation of the present analysis is that it was not based on a written protocol, but merely followed the findings of Naudet and colleagues (18).”

- Then there is no doubt that authors identified differences between placebos. There is no doubt that such a result is counterintuitive. And there is no doubt that this is due to bias. But many other types of bias than unblinding may explain this finding. Authors should therefore discuss in depth this major point in the discussion. Currently the discussion is only supportive of their results and, in my opinion lacks many other perspectives. I rather think that the discussion should emphasize the complexity inherent in interpreting these findings.

**Answer:** We tried to rule out alternative explanations through controlling of covariates. For instance, one common explanation is that the placebo response has increased over time. This is probably true for the period between the 1970s and 1990s, therefore the lower efficacy of amitriptyline and trazodone could be due to the changes in trial participants observed during that period (shift from participants with severe endogenous/melancholic depression to participants with milder forms meeting DSM criteria for major depression). Another important process that took place over time is the narrowing of inclusion criteria and the widening of exclusion criteria, meaning that trial participants became more selective and less representative over time. We agree that with our data at hand we cannot rule out this possibility and thus state this confounding factor in the discussion as an alternative explanation. We now also state in the conclusion section that preregistered confirmatory studies are necessary to test confirm our hypothesis.

**Page 14:** “Finally, it is important to note that our analysis cannot fully rule out alternative explanations. For instance, instead of unblinding, another reason could be the transformation of trial protocols over time. To name just one example, inclusion and exclusion criteria of antidepressant trials have become more restrictive over time, meaning that trial participants are increasingly unrepresentative (44). Although controlling for study year certainly reduces this confounding effect in part, it cannot remove it altogether.”
Page 16: “However, our exploratory post-hoc analysis cannot rule out alternative explanations, which is why the influence of side effects on unblinding should be tested in preregistered confirmatory studies.”

- What about the issue of multiplicity in these numerous comparisons? How many differences are expected to be found on a random basis? By the way, this is a question that the editor might also want to discuss with a statistical reviewer.

Answer: As described in the method section on page 7, the Bayesian analysis used an approach to reduce the problem of multiplicity: “Multiplicity issues were accounted for by using a symmetric random-effects NMA model with exchangeable treatment effects (25), which have been shown to fit well when there is no obvious placebo or other reference treatment in the network, as it was the case in the present analysis.”

In addition, the following paragraph has been added to the limitation section of page 14: “A methodological limitation is the problem of multiplicity. Standard NMA models usually do not account for multiple comparisons in estimating relative treatment effects, which might lead to exaggerated and overconfident statements regarding relative treatment effects. The present analysis therefore applied the Bayesian approximation to reduce that problem described by Efthimiou and White (25), where treatment effects are modelled exchangeable, and hence estimates are shrunk away from large values.”

- Then, it should be recalled that basically, all comparisons between placebos rely of course on indirect evidence (and not on a mix between direct and indirect comparisons). Therefore, the consistency hypothesis (i.e. effects observed between direct and indirect comparisons are the same) cannot be verified. Of course, it would be impossible in this context to verify this hypothesis but we cannot be sure of the validity of the comparisons. This occurs often as even in mixed treatment comparisons; the main part of the evidence is often based on indirect evidence (see 10.1016/j.jclinepi.2020.04.009);

- Indirect comparisons are indeed not robust and very prone to Vibration of effects as we have illustrated elsewhere. And indeed, unblinding of certain studies can be one among many alternative reasons for such a vibration. It is not rare in practice to see overlapping meta-analyses on the same topic with different results. Let’s only focus on the two published version of Cipriani’s meta-analysis.

Answer: The comment of the reviewer has been added to the limitation section on page 14: “Another limitation concerns the evidence summarized in this special placebo NMA, in that all comparisons between placebos rely on indirect evidence only, and not on a mixture of direct and indirect comparisons. Though, in mixed treatment comparisons, the main part of the evidence is often based on indirect evidence (43). The consistency hypothesis, assuming that effects between direct and indirect comparisons are the same, can therefore be verified. Though, it is impossible in this placebo-context to verify this hypothesis, one cannot be sure of the validity of the comparisons considering that indirect comparisons may not be robust and prone to vibration of effects (44).”

- Then the comparisons rely on the similarity hypothesis that assumes that all trials are similar enough to be pooled together. Of course, Cipriani considered this hypothesis to be valid in his study but this is not sufficient to demonstrate that it can be assumed to be true. In my opinion, it is
very challenging to describe and compare all studies of all antidepressants on their major characteristics. In this spirit I really appreciate the numerous adjusted analyses that tried to control for this. But, still some many unmeasured characteristics might have an influence on the authors findings. e.g. outpatients/inpatients or any other surrogate of depression severity at study entry. 

Answer: The comment of the reviewer has been added to the limitation section on page 15: “A more general limitation is that the reliance on the similarity hypothesis that assumes that all trials are similar enough to be pooled together. Cipriani et al. (2) considered this hypothesis to be valid, but still some unmeasured characteristics might have influenced our findings, such as differences between in- and outpatients or any other surrogate of depression severity at study entry.”

- I can easily understand why authors did not provide details on double placebos, that may likely be difficult to interpret. I would like them to make it clear:
  . Was it based on an a priori choice written in a protocol?
  . Why then, these placebos were kept in the analysis?

Answer: The choice to keep double-placebos in the manuscript was to allow readers to have information about the full range of possible placebo comparisons. This way readers are not hold back information that 24% of trials included double-comparisons. The choice to not include the double-placebos in the main analysis was not based on a written protocol.

Again, I think this is a very important and nice paper and even if these limitations can be seen as important, I do not see any alternative design that can address author's questions. I would however expect authors to do the following minor changes:

- Make it clear that the result is only highly exploratory and is important for discussion;

Answer: The term exploratory has been added at various location to the manuscript in the abstract, methods section, results section, and discussion.

- Avoid any spin in the abstract and in the paper (e.g. "Our results indicates that distinguishable side effects of antidepressant may have resulted in unblinding... and consequently in underrating in placebo arms")

- Provide more details about the method in the abstract;

- Provide results in the abstract with point estimates and confidence interval / avoid any interpretation in the results section of the abstract (e.g. Results suggested that...);

- In other words, please use PRISMA abstract to make it better / I suggest to have a long abstract even if the journal requires a small number of words;

Answer: The abstract was changed considering the PRISMA checklist for abstracts, were applicable. More details about methods, results, and limitations have been provided.

“Background: The issue of unblinded outcome-assessors and patients has repeatedly been stressed as a flaw in allegedly double-blind antidepressant trials. Unblinding bias can for example result from a drug’s marked side effects. If such unblinding bias is present for a given drug, then it might be expected that the placebos of that drug are rated significantly less effective than that of other antidepressants.

Methods: To test this hypothesis, the present exploratory analysis conducted a Bayesian network meta-analysis (NMA) comparing the efficacy of 19 different placebos in placebo-controlled trials provided in the dataset by Cipriani et al. (Lancet 2018; 391: 1357–66). Primary outcome was efficacy (continuous) estimated on the standardized mean difference (SMD) scale and defined as
the drug-placebo difference on the Hamilton Depression scale (HAMD-17), on which information was available in N = 258 trials.

Results: Comparative placebo ranking suggested mirtazapine-placebo (SMD -2.0 [-5.0 – 1.0 95% CrI]) to be the most, and amitriptyline- (SMD 1.2 [-1.6 – 3.9 95% CrI]) and trazodone- (SMD 2.1 [-0.9 – 5.2 95% CrI]) placebos to be the least effective placebos. Other placebos suggested to be more effective than amitriptyline- and trazodone-placebos (based on 95% CrIs excluding zero) were citalopram, desvenlafaxine, duloxetine, escitalopram, fluoxetine, sertraline, and venlafaxine placebos. These NMA results were corroborated by the observation that the relative efficacy between drug and placebo was considerably larger for amitriptyline and trazodone than for instance mirtazapine, duloxetine, and venlafaxine, supported by a small and insignificant correlation between drug-efficacy and placebo-efficacy (r = -0.202, p=0.408).

Discussion: The present exploratory NMA indicates that distinguishable side effects of older drugs may have resulted in unblinding of outcome-assessors that may have resulted in an overestimation of the average drug-placebo difference and an underrating bias in placebo-arms, particularly for the older antidepressants drugs amitriptyline and trazodone. These findings suggest that efficacy rankings for antidepressants are susceptible to bias and may considered unreliable or misleading. The analysis is limited by the focus on the single-comparison placebos (76%, i.e., placebos assessed in two-arm trials), since double-comparison placebos (25%, i.e., placebos assessed in three-arm trials) are hard to interpret and therefore not included in the present interpretation. Another limitation is the problem of multiplicity, which was only approximately accounted for in the Bayesian NMA by modelling treatment effects as exchangeable.

Registration and funding: The present analysis was conducted without pre-registration and without formal funding."

- Develop both in the abstract and in the paper the limitations discussed above. I rather expect a strong discussion about uncertainty in this context.
Answer: See the answers to the previous comments regarding abstract and discussion, where limitations have now been addressed in more depth.

- Detail Figure S1 per drug: I would like to know if there are differences across drugs in % of exclusion for instance;
Answer: A new Table S1 has been added to the supplement detailing the excluded drugs due to missing information on efficacy outcome (see below).

- Explain if the fact that mirtazapine placebo = rank 1 was expected a priori? Indeed, to me this appears surprising as mirtazapine (as it carries some anti histaminic properties) is somewhat sedative and therefore, I would have thought that is was an antidepressant very prone to unblinding. This should be discussed in depth.
Answer: The following paragraph has been added to the discussion at page 13 to address the side effect profile of mirtazapine: “For example, mirtazapine, which has a unique dual mode of action as a noradrenergic and specific serotonergic antidepressant (33), has sedating effects due to its affinity to histamine receptors at low plasma concentrations (34). This antihistamine effect, however, is offset at higher doses by increased noradrenergic transmission, which reduces its sedating effect (35-37). Mirtazapine is further considered to have a lower risk of anticholinergic or
- Please explain which effects might be prone to unblinding? Sexual effects? Please also indicate the future studies that one should do to explore/confirm the authors' hypotheses.
Answer: We now stress in the discussion on page that side effects that are easily observable to outcome assessors are presumably those most likely to cause unblinding, e.g. dry mouth, tremor, somnolence. Side effects that are hardly detectable if not reported by the patients, e.g. sexual dysfunction, lack of appetite, would thus be less important. We also propose how a future study could specifically test our hypothesis. We thus added on page 14:
“To confirm our hypothesis, a preregistered prospective study is required. Given that side effects that are observable for an outcome assessor even when not reported by the patient (e.g., dry mouth, tremor, drowsiness, somnolence) are presumably those causing unblinding, it would be worthwhile to examine whether these specific side effects (relative to less detectable side effects such as sexual dysfunction and lack of appetite) lead to correct identification of treatment received and whether they are negatively correlated with depression ratings in the placebo arm.”

- Please justify or avoid the used of somewhat frequentists terms such as "statistically significant" in the context of the Bayesian meta-analysis. This could be assessed too by a statistical reviewer.
Answer: The term ‘statistically significant’ has been remove in all instances in the context of Bayesian analysis and was replaced by the justification of ‘95% credible intervals [CrIs] excluding zero’. For example, in the results section on page 9:
“In accordance with previous re-analyses of the Cipriani dataset, all placebos were less effective than antidepressants (95% credible intervals [CrIs] excluding zero) (1, 3, 4), in line with the main results reported by Cipriani and colleagues (2).“

- It should be clear that all placebos were not active placebo. Can this be checked and explained.
Answer: The following sentence was added to the methods section on page 6: “All placebos were inactive placebos.”

- The comment on spin in the abstract also applies to the conclusion. I would expect it to be way more nuanced.
Answer: According to the abstract, the conclusion has been changed as follows: “Considering clinician-rated symptom change, the present analysis suggests amitriptyline-placebo and in particular trazodone-placebo to be less effective than various SSRI- (citalopram, escitalopram, fluoxetine, sertraline), SNRI- (duloxetine, desvenlafaxine, venlafaxine), and NaSSA- (mirtazapine) placebos. A likely explanation might be that the distinguishable sedative side effects and poorer tolerability of amitriptyline and trazodone may have resulted in unblinding of outcome-assessors and consequently in an overestimation of the of the average drug-placebo difference and an underrating of symptom-change in the placebo-arms. These findings illustrate that efficacy rankings for antidepressants are susceptible to bias and thus may be considered unreliable or even misleading. Unless proven otherwise, it may be assumed that the blind is regularly broken in antidepressant trials when drugs have marked and distinguishable side effect profiles.”
- Then it is very important to discuss in depth the implications of these results not solely for antidepressant research but for the field of evidence synthesis as a whole.
Answer: We added on page 15: “The main implication of our study is that unblinding should be systematically assessed and reported in antidepressant trials. This would allow to statistically control for unblinding effects and it would also be possible to conduct a confirmatory study as detailed above. If our hypothesis holds, it would imply that inert placebos are a poor control and thus the use of active placebos should be reconsidered. Another implication would be that efficacy rankings based on NMA must be interpreted with caution.”

I recommend to accept the manuscript after all these points are taken into account as it is a very nice piece of work that will surely make people think, discuss. It will surely contribute positively to both the field of psychiatry and the field of evidence synthesis.
Answer: The authors thank the reviewer for the helpful and informative comments.

Reviewer 2:
This is an interesting argument and generally well-done study. I have one major concern and two minor suggestions for improvement.
Answer: The authors thank the reviewer for the positive response. All reviewer comments have been addressed and corresponding changes have been made throughout the manuscript

The major concern I have is with the statement in the methods section that their "primary outcome was efficacy (continuous) defined as the drug-placebo difference." However, the abstract, introduction, results, and discussion focus on the effect size for the placebo response, not the drug-placebo difference. What is the reason for this discrepancy?
Answer: The authors agree with the reviewer that the different terminology was misleading. The term 'response' has therefore been changed to 'efficacy' throughout the manuscript.

The authors state that "unblinded outcome-assessors has repeatedly been stressed for years as a major flaw in allegedly double-blind psychiatric drug trials." This problem is not limited to assessors, but also to unblinding of participants. This can be fixed by simply adding the words "and patients" to the sentence.
Answer: The words ‘and patients’ have been added to the sentence.

The authors argue that the nonsignificant correlation between the drug response and the placebo response confirmed their hypothesis that unblinding may have inflated reported drug-placebo differences. They are to be congratulated for avoiding the trap of correlating placebo response scores with drug-placebo differences, as that would be expected as a result of regression towards the mean. However, wouldn't their hypothesis predict a negative correlation between the drug response and the placebo response? This is especially relevant given the findings that breaking blind is more common in drug arms than in placebo arms (e.g., Baethge, Margraf, Rabkin) and that trials in which patients and raters know they are getting an active drug (i.e., trials without placebo controls) produce larger drug responses.
Answer: In our view, and if our hypothesis is true, then the correlation would only be strongly negative if all drugs are easily detectable due to their side effect profile. However, as we state in
the discussion, the newer antidepressants are better tolerated and thus their side effects are less likely to cause unblinding. The most important aspect about this correlation is that both amitriptyline and trazodone clearly deviated from the drug-placebo correlations for other drugs, which is consistent with our hypothesis. That is, both these drugs have good or moderate efficacy, but their placebos show very poor efficacy (strong negative correlation), whereas drugs with less detectable side effects don’t show this characteristic pattern. We clarified in the results section. “Consistent with our hypothesis, the differences in relative efficacy between drug and placebo was considerably larger for AMI/AMIp and TRA/TRAp (that is, negatively correlated) than for instance for MIR/MIRp, DUL/DULp, and VEN/VENp (Fig. 4).”