Title: A Comparison of Depression Prevalence Estimates in Meta-Analyses based on Screening Tools and Rating Scales versus Diagnostic Interviews: A Meta-research Review

Authors:

Brooke Levis (brooke.levis@mail.mcgill.ca)

Xin Wei Yan (xin.yan@mail.mcgill.ca)

Chen He (chen.he3@mail.mcgill.ca)

Ying Sun (ying.sun2@mail.mcgill.ca)

Andrea Benedetti (andrea.benedetti@mcgill.ca)

Brett Thombs (Brett.thombs@mcgill.ca)

Version: 1 Date: 11 Dec 2018

Author’s response to reviews:

Reviewer 1

1. This manuscript describes a meta-review and synthesis of published meta-analyses that document depression prevalence. The review specifically addresses the discrepancies between studies that utilize self-report rating scales, structured diagnostic interviews, and a combination of these methods. The topic is of great importance to researchers and practitioners who are concerned with depression and the methods utilized to inform study conclusions are well-done and appropriate. However, the authors are notably dismissive of rating scales and their established validity while largely ignoring a number of well-documented shortcomings in structured diagnostic interviews. The following specific critiques and suggestions are offered:

We thank Reviewer 1 for his appreciation of the importance of the topic of our study and the recognition that the methods are well done and appropriate. We also thank Reviewer 1 for pointing out that we did not clarify for readers in our original manuscript that commonly used depression symptom questionnaires and rating scales are, indeed, validated as screening tools with documented sensitivity and specificity, but not to identify cases of depression or to estimate prevalence. We also appreciate his comment that while validated diagnostic interviews have been validated for identifying cases and establishing prevalence, they have some shortcomings. We hope that our edits have clarified this.
The title of the manuscript suggests the unsupported conclusion that higher estimates produced by studies that utilize rating scales, when compared to those relying on structured diagnostic interviews, necessarily represents overestimation.

We thank Reviewer 1 for pointing out that we did not adequately explain this. Published, validated cutoffs on commonly used depression screening tools are validated for the purpose of screening. That is, sensitivity and specificity of the tool for depression detection have been assessed, but these tools are not validated for the diagnosis of cases or to estimate how many patients in a population would have depression. Validated diagnostic interviews, on the other hand, albeit imperfect, are validated for precisely the purpose of diagnosis and, thus, establishing prevalence. To address this point, we have added Table 1. Table 1 shows, based on estimates of sensitivity and specificity for several commonly used depression screening tools from meta-analyses, the degree that each would overestimate prevalence if used for that purpose. The table shows the extent of overestimation both in raw terms and as a ratio. In addition, we have added text to the introduction to explain this more clearly than we did previously (Lines 77-99).

Administration of validated diagnostic interviews is time and resource intensive, however. Thus, instead of validated diagnostic interviews, researchers sometimes use self-report depression symptom questionnaires, or screening tools, and report the percentage of patients above standard screening cutoff thresholds as prevalence [8]. These screening tools and standard cutoffs, however, have been validated and calibrated to maximize sensitivity and specificity for screening, but not for assessing prevalence. Screening cutoffs are typically set to cast a wide net and identify many more patients than the number who meet diagnostic criteria. It would be possible to set cutoffs on self-report depression symptom questionnaires to estimate prevalence in a population, rather than for screening [8, 9], but we do not know of any examples of self-report tools with cutoffs calibrated to estimate the prevalence of major depression or any other psychiatric disorder.

The degree to which the percentage of patients scoring above a screening tool cutoff inflates prevalence depends on the specific screening tool and cutoff used, but most commonly used screening tool and cutoff combinations exaggerate depression substantially. Because the false positive rate of screening tools is disproportionately high in lower-prevalence populations, such as primary health care, estimated prevalence based on screening tools is exaggerated most when true prevalence is lowest [8]. The expected percentage of patients who score above a cutoff on a depression screening tool can be calculated from the tool’s sensitivity and specificity and the true population prevalence [8]. Table 1 shows the percentage of patients who would score above standard cutoffs for screening for commonly used depression screening tools based on sensitivity and specificity estimates from meta-analyses and for true prevalence of 5%, 10%, and 15%. The degree of overestimation of true prevalence varies across screening tools and cutoffs, but it is substantial in almost all cases.

We have also changed the title, per Reviewer 1’s suggestion to read, “Misuse of Screening Tools or Rating Scales Instead of Validated Diagnostic Interviews to Estimate Depression Prevalence in Meta-analyses: A Meta-research Review”
3. The use of the term "validated" is applied only to structured diagnostic interviews in this context, when both types of case ascertainment method are, in fact, generally validated as a precondition of their use and publication.

Screening tools and cutoffs have been validated and calibrated for the purpose of screening, and this is typically done by jointly maximizing sensitivity and specificity. This is not the same, however, as being validated to diagnose cases or to estimate prevalence. In the revised version, we explain this more clearly (Lines 77-87):

Administration of validated diagnostic interviews is time and resource intensive, however. Thus, instead of validated diagnostic interviews, researchers sometimes use self-report depression symptom questionnaires, or screening tools, and report the percentage of patients above standard screening cutoff thresholds as prevalence [8]. These screening tools and standard cutoffs, however, have been validated and calibrated to maximize sensitivity and specificity for screening, but not for assessing prevalence. Screening cutoffs are typically set to cast a wide net and identify many more patients than the number who meet diagnostic criteria. It would be possible to set cutoffs on self-report depression symptom questionnaires to estimate prevalence in a population, rather than for screening [8, 9], but we do not know of any examples of self-report tools with cutoffs calibrated to estimate the prevalence of major depression or any other psychiatric disorder.

We also provide guidance in the discussion on how researchers could calibrate cutoffs that would be validated for estimating prevalence (Lines 350-356):

Another approach, prevalence matching, would involve calibrating cutoffs on depression symptom questionnaires to estimate case prevalence in a population rather than to maximize sensitivity and specificity for screening. This could be done by administering a screening tool and a validated diagnostic interview to all patients in a study and setting a cutoff score that results in the percentage above the cutoff matching as closely as possible the number of patients with depression, based on the validated diagnostic interview [8, 9]. We do not know, however, of any examples where this has been done.

4. Ln74-80: The authors describe research that utilizes structured diagnostic interviews as adhering to a "gold standard." This is problematic for a few reasons. Most studies that use structured diagnostic interviews to report case "diagnosis" rates rely on appropriately trained students or lay community members to conduct these interviews. This removes interview results from the context in which clinical diagnosis is made. A number of studies (see Rettew et al., 2009, Eaton et al., 2000, and others) show that when compared to practicing clinicians, typical administrations of structured diagnostic interviews show poor agreement and may underestimate diagnosis in certain groups. While this suggests against the use of absolute language like "gold standard," a reasonable alternative would be regarding structured diagnostic interviews as a case ascertainment method with its own strengths and weaknesses. This might create an opportunity for a more productive discussion of what can be learned in the comparison of findings from these different methods.
Indeed, even validated diagnostic interviews have shortcomings. We have removed the term “gold standard”. Also, in the discussion, we provide references to studies that have shown that different diagnostic interviews seem to perform differently and be differentially reliable, and that they may not always be applied as intended, as pointed out by Reviewer 1 (Lines 377-383):

A second possible limitation is that we did not examine the differential performance of different types of validated diagnostic interviews as this was beyond the scope of the study. Indeed, there are differences in the performance of different validated diagnostic interviews, as we demonstrated in a recent meta-analysis [36], and different types of diagnostic interviews may be differentially reliable [37, 38]. Furthermore, these interviews may not be used in the way that they are intended or by the types of interviewers for who they are designed, which could also influence their performance.

5. Ln89: Related to the previous concern, the assumption that rating scales necessarily inflate prevalence seems to be an overreach.

Cutoff scores on depression screening tools would not overestimate or inflate depression prevalence if the cutoffs were calibrated to estimate prevalence. We are not aware, however, of any examples of that. Rather, published cutoffs have been set to maximize sensitivity and specificity. We hope that our revised manuscript explains that this does, mathematically, lead to inflated estimates of prevalence if the percent with positive screens is misreported as prevalence. As an analogy, reporting positive screens as cases would be akin to reporting the number of women with positive mammography screens as the prevalence of breast cancer. We have added Table 1 to clearly show this and provided a clearer explanation in the text (Lines 88-99):

The degree to which the percentage of patients scoring above a screening tool cutoff inflates prevalence depends on the specific screening tool and cutoff used, but most commonly used screening tool and cutoff combinations exaggerate depression substantially. Because the false positive rate of screening tools is disproportionately high in lower-prevalence populations, such as primary health care, estimated prevalence based on screening tools is exaggerated most when true prevalence is lowest [8]. The expected percentage of patients who score above a cutoff on a depression screening tool can be calculated from the tool’s sensitivity and specificity and the true population prevalence [8]. Table 1 shows the percentage of patients who would score above standard cutoffs for screening for commonly used depression screening tools based on sensitivity and specificity estimates from meta-analyses and for true prevalence of 5%, 10%, and 15%. The degree of overestimation of true prevalence varies across screening tools and cutoffs, but it is substantial in almost all cases.

6. Ln100: The assumption that rating scales necessarily overestimate is repeated.

Please see our response to the previous comment.
7. The approach to aggregating and summarizing study findings is well done and appropriate.

We thank Reviewer 1 for recognizing this.

8. Ln224: It would be very interesting to see the proportion of studies that utilized structured diagnostic interviews administered by practicing clinicians. If these are sufficient in number, this could be a focus of sensitivity or subgroup analysis.

We agree that validated diagnostic interviews, whether they be semi-structured or structured, have flaws. Indeed, there are few examples of diagnostic procedures, whether they be for cancers, mental health disorders, or other conditions, that are not imperfect. We hope that we have clarified in the revised manuscript the importance of using methods intended to classify cases to establish prevalence and that screening cutoff scores are not intended nor validated for this purpose and would, mathematically, lead to rates that do not reflect prevalence. We have addressed this in other comments. We have also addressed the idea that diagnostic interviews are not perfect classification methods in the section on limitations, but have noted that an examination of the differential performance of differential interviews is out of the scope of the research questions we asked and beyond our methods (Lines 377-383):

A second possible limitation is that we did not examine the differential performance of different types of validated diagnostic interviews as this was beyond the scope of the study. Indeed, there are differences in the performance of different validated diagnostic interviews, as we demonstrated in a recent meta-analysis [36], and different types of diagnostic interviews may be differentially reliable [37, 38]. Furthermore, these interviews may not be used in the way that they are intended or by the types of interviewers for who they are designed, which could also influence their performance.

9. A more appropriate title and tone for this manuscript would omit foregone conclusions about one measurement approach or the other and instead directly address what occurs in the methods: two popular approaches to case ascertainment that are widely reported in meta-analyses are quantitatively summarized and compared.

As we indicated in our response to Comment #2, we have changed the title to reflect this concern. Our revised title is “Misuse of Screening Tools or Rating Scales Instead of Validated Diagnostic Interviews to Estimate Depression Prevalence in Meta-analyses: A Meta-research Review.” As we described in our responses to Comments 2, 3, and 5, it is the case that when cutoffs set to maximize sensitivity and specificity for screening are used to estimate prevalence, that prevalence will be, in almost all cases, overestimated or inflated. This is discussed in more detail in the revised version in the introduction and shown in Table 1. Thus, this is a mathematical reality and not speculative. In the discussion, we also describe that depression symptom questionnaires or screening tools could be validated for estimating prevalence, but that would require specific studies for that purpose, and we do not know of any such studies. As we have described in the manuscript, using cutoffs designed for screening to estimate prevalence
does not have evidence to support this and has not been validated. Rather, based on existing evidence, we know that it is not valid.

Reviewer 2

1. I only have a few albeit major comments.

We thank Reviewer 2 for his helpful feedback regarding our manuscript. We hope that our responses below and our revisions to the manuscript have adequately addressed the comments that were made.

2. I know the authors state in the Discussion that their goal was not to compare actual magnitude of prevalence rates. However, the very basis of their overall conclusions is that these are different depending on the assessment method (objective 2). When I read through the description of the primary studies included, I noticed that many different populations (also age) and settings were included. Just for example, in Table 1a, one MA included children and adolescents in the welfare system, and another one Post-stroke patients in Sub-saharan Africa. Overall, the primary studies (MAs) seem to come from very different settings and it is not clear whether it was appropriate to compare their prevalence to begin with or what effect these different settings had on the differences in pooled prevalence rates.

We agree that it is not appropriate to compare the prevalence across populations. We have tried to emphasize more clearly in our revised manuscript that using cutoffs designed for screening to estimate prevalence would, mathematically, based on what we know about sensitivity and specificity for these tools, lead to inflated prevalence. We note the amount of heterogeneity in the study limitations (Lines 363-371), and we also note that there is a need for future work to quantify the extent of inflation based on specific screening tools, since this can be quantified (Lines 374-376).

3. In the same line, please provide forest plots to get a sense of the variation of the prevalence rates across MAs, in particular for the mean differences reported for objective 2 (I assume these mean differences are within comparisons, please correct me if I am wrong). Not weighting them seems appropriate in the given context, but a stronger case could be made for doing so. Perhaps a sensitivity analysis using weights could be considered.

Objective 2 is mostly for “across-methods” comparisons as a whole – only 12 studies have within-study comparisons. As we describe, when there were direct comparisons, the prevalence was on average 10 percentage points higher with screening tools (Lines 243-246). We do not believe that weighting would be appropriate given the high degree of heterogeneity. To better illustrate the findings, as suggested, we have generated forest plots of pooled prevalence
estimates (and CIs) for each category, and these are reported in Additional Figures 1a-1c in the Appendix, as described on Lines 239-240.