PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Psychological interventions for chronic non-specific low back pain: protocol of a systematic review with network meta-analysis
AUTHORS	Ho, Emma; Ferreira, Manuela; Chen, Lingxiao; Simic, Milena; Ashton-James, Claire; Comachio, Josielli; Hayden, Jill; Ferreira, Paulo

VERSION 1 – REVIEW

REVIEWER	Cary Reid Weill Cornell Medicine New York, NY USA
REVIEW RETURNED	19-Nov-201

GENERAL COMMENTS	This is a well written protocol manuscript that describes a proposed systematic review and meta-analysis evaluating psychological therapies for the management of chronic back pain. Given increasing emphasis on the use of behavioral approaches to manage pain, such a review is timely and will be of great interest to practitioners and researchers. Major concerns include 1. It is not clear how the authors will categorize the various psychological interventions since many of them include combinations of the approaches the authors list on page 6 (e.g., health coaching + CBT), some a priori description of how they will categorize therapies that contain multiple components would be important to do. 2. There is no attention to process variables, e.g., such as number of sessions attended, adherence with homework exercises which can impact treatment outcomes. 3. What criteria will the investigators employ around whether sufficient number of studies are available for a given outcome to conduct a meta-analysis for that outcome? 4. The investigators provide no rationale for the outcomes selected (either as primary or secondary). Interesting that no functional measures appear to be listed. Other issues 1. Abstract page 2, line 9. Would add "patient with" before chronic low back pain in the sentence.

REVIEWER	Matthew K Bagg Neuroscience Research Australia and Prince of Wales Clinical School, University of New South Wales, Sydney, Australia Dr Hayden and I co-authored a Letter-to-the-Editor commenting on a recent network meta-analysis (modes of exercise training for low back pain) in BJSM. Drs M Ferreira, P Ferreira, Hayden and I collaborated to deliver a workshop on network meta-analysis methodology at the 2019 Back

	and Neck Pain Forum in Quebec City. I have no current collaborations with any of the named Authors.
REVIEW RETURNED	23-Dec-2019

GENERAL COMMENTS	**Introductory statement** Thank you for the opportunity to review this manuscript. The paper describes the protocol for a systematic review and planned network meta-analysis to determine the comparative effectiveness of psychological interventions on function and pain, among other outcomes, for people with chronic low back pain. The review will answer an important clinical question and specifies patient-relevant outcomes. I have not found any evidence of similar reviews published, or registered elsewhere. I commend the Authors for their transparency in producing a protocol. I have responded to each of the questions on the Review Checklist and included additional comments to the Authors throughout. **Review Checklist** Mark each as Yes or No or N/A Please elaborate on any 'No' answers in the free text section below. ***1. Is the research question or study objective clearly defined?*** *No* I think the latter half of the Introduction could be re-structured, which would define the research question more clearly. Currently, the Authors position the use of psychological interventions within clinical care for low back pain, before making the statement (that I believe underpins the research question) that "…it remains unclear which psychological interventions offer better benefits for managing chronic LBP,…". This is followed by an exposition of the evidence available for clinical decision-making in previous reviews and the implication of the nature of this evidence for decision-making, a description of network meta-analysis (NMA), and finally, a relation of the capability of NMA to the aforementioned problems. I lost track of the statement underpinning the research question whilst reading this. Consequentially, I didn't clearly understand why the proposed methodology was needed, nor what the research question was. I think the exposition of the available evidence and implications for decision-making should come first, followed by the statement that underlies the research question. I strongly suggest a separate paragraph to describe NMA, before

concluding the Introduction with the Aims. I think this will give the reader opportunity to consolidate the statement underpinning the research question before being introduced to the methodology. The research question(s), espoused in the Aims, are then a logical progression from problem : solution : Aims.

I have the following additional comments to the Authors regarding the Introduction:

Pg 5, Line 42-51
I do not think the lengthy sentence describing the *Lancet* Series working group is necessary. It appears to be trading on the name of the *Lancet*. I suggest deleting that sentence and revising the preceding sentence. For example, consider: "International clinical guidelines and LBP research experts endorse the integration of psychological interventions in the management of chronic LBP.[15-20]"

Pg 5, Line 54
I think it may assist your argumentif you describe the number of different interventions available, from which clinicians must make a decision. Please consider this?

Pg 5, Line 59
Is it only the interventions that are commonly used clinically that have been subject to systematic review? What data do you have describing clinical practice patterns? I think it would assist your argument to describe for the reader what these interventions are and how they are used. To my mind, this justifies their inclusion in your study, irrespective of the previous reviews. If they are neglected in previous reviews this substantiates their inclusion.

Pg 6, Line 3
What do you mean by 'vigorous'?

Pg 6, Line 9-10
You state that "… uncertainty surrounds which approaches policy makers should recommend to clinicians for managing chronic LBP." I think that uncertainty in selecting an intervention as a result of lack of evidence presumes that decision-makers perceive i) a lack of evidence, or ii) that research evidence is needed to make this decision. Do you know if this is the case for this clinical decision scenario? Secondly, do you think this uncertainty applies only to policy makers?

Pg 6, Line 10
The correct spelling is 'undoubtedly'.

Pg 6, Line 11-13

	How do you know that uncertainty in the research evidence creates a lack of confidence in clinicians? I don't think the reviews you have cited substantiate this claim. My reading of these two reviews is that whilst they do state that physiotherapists have a perceived lack of competence in managing these factors, they do not mention physiotherapists' use of evidence to select interventions, let alone state that they have a lack of confidence in managing these factors. I interpret this as evidence contrary to your claim. I think it suggests that clinical decision-making may occur without reference to research evidence, which implies that the state of the evidence does not influence clinicians' confidence. I only know of a single other review relevant to the topic of physiotherapists' decision-making: Gardner et al. 2017 *J Physio*. They focused specifically on beliefs and attitudes (admittedly, this does not include knowledge, which may relate more closely to research evidence) that influence the selection of interventions. I can find no statement that this is done using research evidence. I think the claim that uncertainty of evidence about relative effectiveness may make it more difficult to make decisions is reasonable to make, however we need to acknowledge the apparent lack of data substantiating this. Your claim seems to me to be reliant on the presumption that clinicians seek to use research evidence in their decision-making. I think your argument will benefit from sourcing evidence to substantiate the claim, or from expressing the claim with due acknowledgement to the data. *Pg 6, Line 13 and Line 17* I think it will strengthen your argument for the use of NMA if you describe the assumptions upon which 'robust' network meta-analytic inference is based. Whilst not your intent, your current phrasing implies that NMA is robust by default. It is important that readers are aware that NMA must be conducted with appropriate consideration of key assumptions, namely transitivity and coherence (consistency). Otherwise, causal interpretation of the inference is not valid and the methodology provides no advance on comparing across pair-wise analyses. *Pg 6, Line 16-17* I think you are saying too much in a single sentence here and are obscuring meaning. I would rephrase in view of the fact that producton of rankings is a separate, ancillary step to production of effect estimates. The comparison of interventions (in the NMA model) produces the estimates of relative effectiveness. Those estimates are then used to generate rankings. *Pg 6, Line 17*

	Please clarify the nature of the inference? Does the model compare study arms, or interventions? *Pg 6, Line 21-22* I don't think this is correct. Indirect estimates of effect may be derived for *any* comparison between interventions that share a common comparator. These indirect or mixed comparisons may be of newer interventions, older interventions, active-active or active-sham. Whilst it is usual for active-active comparisons to be evaluated comparatively less than active-sham comparisons for interventions that require regulatory approval (e.g. medicines), I am not sure this is true of psych interventions for LBP? I would keep this statement general, unless you have an idea of the structure of the evidence-base from prior work? *Pg 6, Line 28* In line with my previous points, I do not think it is appropriate to deem the evidence we produce crucial to decision-making, when the data suggest this evidence may not be used. *Pg 6, Line 29* Please clarify what you mean by 'better benefits' - better than what? Has it been established that *all* psychological interventions for LBP are beneficial? ***2. Is the abstract accurate, balanced and complete?*** *Yes* ***3. Is the study design appropriate to answer the research question?*** *Yes* The research question is: "what is the most effective psychological intervention for chronic low back pain?", although it could be worded more clearly. A SR and NMA is the appropriate design to answer this question. ***4. Are the methods described sufficiently to allow the study to be repeated?*** *No* **The following items need to be clarified, please?** *Pg 6, Line 49* How will you account for the clustering amongst observations in cluster-randomised trials? They cannot be used in a NMA unless this has been adjusted for *Pg 6, Line 50* Your statement that you will include '… any comparison interventions' is misleading, in my view, because you later

	define the set of interventions. '… any comparison interventions' implies you will include *all* interventions. Please clarify? *Pg 6, Line 51* What will you do with cross-over studies that have more than two phases? *Pg 6, Line 56* Why are you not considering unpublished data sources? *Pg 6, Line 56* Do you mean that you will include only those records for which you are able to retrieve a full-text? Or that you will only consider published articles that are full-length journal articles? What about short reports, research letters and conference abstracts? *Pg 7, Line 5* What about somatic leg pain? Do you consider either somatic or radicular leg pain an effect modifier? *Pg 7, Line 18-19* Where in Neilson & Weir (2001) is this definition? I can't find it. I am concerned that this definition is insufficient for your purpose here. Do you mean to say that *any* clinical interaction that influences the psychological experience is an intervention of interest for this review? By that reasoning, what interventions do we not include? What will you do with interventions identified during the review that meet these criteria, but are not listed here? I strongly encourage you to clarify this definition to provide reasonable: 1. justification for the list of interventions as it stands 2. guidance for including interventions identified during the review that are not listed here *Pg 7, Line 24* What about other forms of education? Will you include reassurance? What forms of biofeedback? (This may become irrelevant once you have clarified the interventions of interest) *Pg 7, Line 26-27* 'direct comparisons between different types of psychological interventions' is not an intervention *Pg 7, Line 29* How do you define exercise regimina? Do you consider physical activity exercise? *Pg 7, Line 31*

	How do you define ‘conservative management’? *Pg 7, Line 46-50* What will you do with studies that don’t provide a NRS or VAS, yet measure pain intensity? *Pg 8, Line 20-24* What platform will you use to search these databases? *Pg 8, Line 34-41* Please clarify your intent here? You are not including unpublished reports of randomised trials? So it would appear you are searching registries to identify additional data? If so, how will you link the multiple records for each trial and prioritise the use of data from these records? What will you do when data sources conflict? If you are only searching registries to identify registered trials that havbeen published, why bother? This is unlikely to retrieve published trials that were not retrieved by database searches because the linkage of published records to trial registrations is imperfect. And why not include unpublished records, given that you will screen them? *Pg 8, Line 46* I think you mean that you will screen records in Covidence. What citation manager will you use to manage the search output? How will you manage records from trial registries? *Pg 8, Line 46-50* It appears that you will not screen studies in two stages? *Pg 8, Line 58* Will this form be piloted? Has it been used before? *Pg 9, Line 7* Will you extract data at the study-level or the group-level? What do you mean by drop-out? What are the measures of central tendency, dispersion or frequency that you will extract? For what time points will you extract data? *Pg 9, Line 14* What are these data? Contrast-level or group-level? Point estimates and measures of precision, or measures of dispersion? What will you do if you can’t extract SDs for continuous outcomes? *Pg 9, Line 19* What criteria determine whether data from a trial are included in the analysis? Must the intervention be complete? Are these time points with respect to randomisation or end of

	treatment? Do you think it is reasonable to group outcomes in these broad bands? *Pg 9, Line 22-38* It appears you will categorise nodes based on clinical appropriateness as well as the available data? I think there is a risk of your judgements being influenced by the data. It would be better (and in-line with guidance) to describe the criteria for node definition in a separate paragraph when you define the interventions. These criteria should be relevant factors for the clinical question, e.g. dose, setting, care-provider and have an empirical or strong theoretical justification for an association with the outcome. *Pg 9, Line 27* You are including medicines as well? This is why the set of interventions of interest needs to be more clearly defined. As above, it is not clear what Ixs you are including. *Pg 9, 29-33* What will you do with combination interventions otherwise? *Pg 9, Line 33-34* I don't understand the difference in meaning between 'distinct' and 'single'? *Pg 9, Line 36-38* Great, although I don't know what your proposed network geometry is - the network of all possible comparisons that should be described at protocol stage. *Pg 9, Line 55-59* How will you make the overall risk of bias judgement? *Pg 10, Line 53* What is your justification for these covariates being effect modifiers? Do you consider leg pain an effect modifier? *Pg 10; Line 57* What criteria will you use to make a judgement of whether there is insufficient transitivity? Please note that transitivity is not a binary construct. Is this pair-wise or network meta-regression? *Pg 12, Line 3-10* I strongly encourage you to check these statements. CINeMA is a framework for applying the GRADE approach to form a judgement of confidence (not quality) in the evidence. There are some Cochrane training modules for CINeMA, however it was developed by Georgia Salanti's team at ISPM, Uni Bern.

	The following items need to be included, please? *Types of Studies* What will you do with trials that evaluate some interventions that you are interested in and some that do not? Will you exclude the study, or include the study and extract only the comparisons of interest? *Types of interventions* I think a clear definition of the set of interventions of interest for this review will improve this section considerably. Please also consider defining a decision subset and a comparison subset? What will you do with combination therapies? These are therapies that contain 2 or more interventions, where at least one is of interest to the review. *Assessment of transitivity assumption* How will you assess the similarity of these covariates across comparisons? ***5. Are research ethics (e.g. participant consent, ethics approval) addressed appropriately?*** *Yes* ***6. Are the outcomes clearly defined?*** *No* The outcomes are missing the definition of the construct that is being measured. Consequentially, it is not clear what, for example, form of disability is being referred to. I infer from the outcome measures that this is low back-specific disability. I encourage the authors to re-phrase this as back-specific function and to define each outcome using: domain, construct, measure(s), time point(s). Disability might thus be written as 'Function, defined as low back-specific function, measured using the ODI, …., at xxxxx (time point). I have these additional comments for the Authors: Please justify your choice of the time points for analysis? It is important that these are set at a time point at which it is reasonable for change in outcome under *each* intervention to have occurred. I am not experienced with all of these interventins. Do you think it is reasonable for any hypothetical participant to have experienced a change in outcome under any one of these interventions by each of the time points that you specify? I also encourage you to place the outcomes in a stand-alone section immediately after 'Study Design'. I know it is convention in SRs to write the outcomes for the review as part of the PICOS. However, I think this is confusing because these are

	not inclusion criteria. Thirdly, I would move the description of the hierarchy of measures for data extraction to the Data Extraction section, as this is not related to the definition of the outcomes. Lastly, I am interested in your rationale for not evaluating safety in some manner? Even acceptability of therapy (perhaps defined as all cause discontinuation during the intervention period) would be helpful? ***7. If statistics are used are they appropriate and described fully?*** *No* **The following items need to be clarified, please?** *Pg 10, Line 3-14* What criteria will you use to determine whether the data are appropriate for meta-analysis? Will you reproduce the NMA in R, or will you use STATA and R for different things? What packages will you use? Will you fit fixed or random effects models? What is your justification? What are your assumptions for the network heterogeneity variance parameter? *Pg 10, Line 17-24* Are these the summary measures of effect for the analysis? Please use a more descriptive sub-heading? What constitutes a sufficiently different rating scale? Will you transform any of the scales? How will you do a NMA on a particular outcome if you deem that some studies have used the same rating scale and others have not? Will you use SMD in all cases? If you use SMD, are you aware of the influence of the assumption of common variance on the interpretation of the results? *Pg 10, Line 31* I think it is clearer to say that you will weight the edges proportional to the number of studies evaluating that comparison. *Pg 10, Line 43* I2 does not quantify heterogeneity and it is not a threshold (see Borenstein et al. 2017 *Res Synth Meth*). Please re-phrase this? *Pg 10, Line 48* What are your criteria for insufficient data for a pairwise meta-analysis?

	Pg 11, Line 7 What are the conditions required for preserving within-study randomisation? How will youformulate and assess the transitivity (and other) assumption(s) to meet these requirements? *Pg 11, Line 7-10* I would consider effect sizes for the comparisons between interventions to be relative. Whereas, rankings are absolute. How will you calculate the 95% CI for the SUCRA, given that this is an AUC statistic? *Pg 11, Line 27-29* This will likely also change network structure. How will you interpret this? *Pg 11, Line 29-43* Are you aware that meaningful subgroup inference in a NMA is contingent on the network structure being the same for each subgroup? What will you do if the subgroups have different structures? **The following items need to be included, please?** *Pg 10, Line 27* Do you have criteria for how connected the network must be to proceed with analysis? *Pg 10, Line 43* Will you test for true-study variance in pairwise comparisons using Cochran's Q test? Will you estimate the heterogeneity variance parameter for pairwise comparisons? *Pg 10, Line 57* What are your assumptions regarding the regression co-efficients, the nature of the interaction and the direction of the effect of each covariate? Will you fit a single or multiple covariate(s) in each model? Will you fit these models in a frequentist or a Bayesian framework? *Pg 11, Line 14* How will you assess network heterogeneity? Node-splitting is more appropriately used as a test of inconsistency to supplement the GDBTIM and the loop-specific approach, given all approaches have low power. Have you considered methods that account for studies with >2 groups? What are the criteria by which you will judge consistency? What will you do if you detect inconsistency? ***8. Are the references up-to-date and appropriate?***

The references are up-to-date. However, there appear to be minor capitalisation errors in the names of most of the journals. This is probably a citation software feature that is easily fixed.

9. Do the results address the research question or objective?

N/A

10. Are they presented clearly?

N/A

11. Are the discussion and conclusions justified by the results

N/A

12. Are the study limitations discussed adequately?

No
They will be satisfactorily addressed if the authors respond to my comments regarding Methods and Statistical Analyses.

13. Is the supplementary reporting complete (e.g. trial registration; funding details; CONSORT, STROBE or PRISMA checklist)?

The authors have prospectively registered their systematic review on PROSPERO, and provided the ID number in accordance with *BMJ Open* policy.
The authors have also provided a PRISMA (2009) checklist, which has been accurately completed. However, the authors have not provided two further relevant reporting standards:
1. PRISMA-Protocols (Shamseer et al. 2015 *BMJ*)
2. PRISMA-NMA (Hutton et al. 2015 *Annals Int Med*)

I think the authors should use the relevant items from these standards in the revision of their paper. I also recommend Chaimani et al. 2017 *J Clin Epi* as a great resource expounding the additional considerations required for a planned network, compared to a pair-wise, meta-analysis.

14. To the best of your knowledge is the paper free from concerns over publication ethics (e.g. plagiarism, redundant publication, undeclared conflicts of interest)?

Yes

15. Is the standard of written English acceptable for publication?

Yes

Statistics

1. Does this paper require specialist statistical review?

	Yes and I have performed this review Would you be willing to review a revision of this manuscript? *Yes* ***Recommendation*** *Major Revision*

REVIEWER	Roger Hilfiker School of Health Sciences, HES-SO Valais-Wallis, Switzerland
REVIEW RETURNED	03-Jan-2020

GENERAL COMMENTS	The manuscript of the protocol entitled “Psychological interventions for chronic non-specific low back pain: protocol of a systematic review with network meta-analysis” is well written and the planned systematic review with a pairwise and a network-meta-analysis will be relevant. I have some minor comments, mainly to the method section: Please consider to use the PRISMA Network Meta-Analysis Extension https://annals.org/aim/fullarticle/2299856 Page 6, line 38 to 45: please be more specific for the proposed hierarchy, how will you decide on the order if none of the first four disability measures are present (and more than one is presented)? Page 6, line 47 to 50: please be more specific for the proposed hierarchy: what will you choose if no NRS but several pain rating scales are present? What will you do if different time-points (pain last 3 months, last week, etc.) are present. What will you do if different pain intensity scales are present regarding average pain, worst pain etc. Page 6 line 52 to page 7 line 5: Please be more specific for the proposed hierarchy: What do you do if only subscore (e.g. sf-36 subcomponents) are provided and no overall-score? Page 6, line 16 to 31, (Types of interventions and comparators) and page 10, line 14 to 22: Will you consider to split interventions if inconsistency is present? See James, A., Yavchitz, A., Ravaud, P., & Boutron, I. (2018). Node-making process in network meta-analysis of nonpharmacological treatment are poorly reported. Journal of clinical epidemiology, 97, 95-102. And also : Caldwell, D. M., & Welton, N. J. (2016). Approaches for synthesising complex mental health interventions in meta-analysis. Evidence-based mental health, 19(1), 16-21. How do you deal with doses of the interventions? Page 7 line 7 to 14: Please be more specific, what will you do if only subscore of e.g. FABQ is provided? Page 8, line 15 to 19, please provide mutually exclusive categories: “Data will be classified according to short-term (<6 months), mid-term (6-12 months) or long-term follow-up (≥12 months)” Page 8, line 55 to 59, please provide the rule on how to produce the overall judgment based on the five domain-level judgment Page 9, Data Analysis: could you please be more specific about the approach used within Stata and R (i.e. which ado or package will

	you use). Page 9, Pairwise meta-analysis (Line 38 to 48): please consider to mention why you would not use the Hartung-Knapp-Sidik-Jonkman approach. Page 9, summary measures: if you expect dichotomous data, please specify this in the “types of outcome measures” Page 10, network meta-analysis: please consider to add the reference for frequentist “sucra” values Rücker, G., & Schwarzer, G. (2015). Ranking treatments in frequentist network meta-analysis works without resampling methods. BMC medical research methodology, 15(1), 58. And please consider to indicate that your SUCRA is the p-score which is the frequentist analogy to the baesian SUCRA. (see reference Rücker & Schwarzer 2015). Page 10, assessment of inconsistency: please be more specific on how you implement the Bucher method (which software, which Stata ado or R package). Will the publication bias also be assessed within the network-meta-analysis or only in the pairwise meta-analysis?

VERSION 1 – AUTHOR RESPONSE

Reviewer #1

1. *It is not clear how the authors will categorize the various psychological interventions since many of them include combinations of the approaches the authors list on page 6 (e.g., health coaching + CBT), some a priori description of how they will categorize therapies that contain multiple components would be important to do.*

Author’s response: We have included an a priori description of how we will categorise the various psychological interventions. First, we will categorise the psychological interventions according to the following categories/treatment nodes: (1) “*behavioural therapies”*; (2) *“cognitive behavioural therapies”;* (3) *“mindfulness-based therapies”;* (4) *“counselling-based therapies”;* (5) *“pain education”.* We have also provided a detailed justification and explanation of our classification system in our revised manuscript, including a table (p.4-6).

Regarding interventions that include a combination of psychological approaches, we will also form a separate treatment node *“combination of psychological interventions.”* This node will include interventions consisting of more than one type of psychological approaches that span across the pre-specified categories (e.g. health coaching (i.e. counselling-based therapies) combined with cognitive behavioural therapy). As described by Caldwell et al. (2016), our decision to lump combination interventions into one node is for practical reasons, as we anticipate few eligible studies.[1] If however, we find that there are sufficient studies in distinct combinations of the pre-specified

psychological approaches, we may consider splitting the nodes. This will be decided by consensus amongst reviewers (p.13).

After categorising the psychological interventions into the above categories, we will then further differentiate any non-psychological co-interventions included if present. Non-

psychological co-interventions will be classified as exercise or passive (definitions included in the revised manuscript). As an example of the category system, we will include a separate treatment node labelled as "*cognitive behavioural therapies plus exercise."* We have chosen to differentiate the non-psychological co-interventions as we anticipate that the psychological interventions will be heterogenous for many reasons, including dosage and the psychological strategies used. However, the presence of non-psychological co-interventions is also common, and previous relevant reviews have not accounted for this. This may have contributed to increased heterogeneity and possibly inaccurate estimates of treatment effects (p.13).

2. *There is no attention to process variables, e.g., such as number of sessions attended, adherence with homework exercises which can impact treatment outcomes.*

Author's response: We will extract all data reported on the dosage and frequency of the interventions. We will extract all available data on adherence, including the authors' definition of adherence, and number of sessions attended (p.11). Process variables are not commonly reported well or consistently in studies of psychological interventions. Hence, subject to availability of data, we will attempt to perform meta-regressions based on intervention parameters relating to dosage and/or frequency, for example, by total length (in weeks) of the intervention or total intended hours of the intervention during the intervention period (p.18).

3. *What criteria will the investigators employ around whether sufficient number of studies are available for a given outcome to conduct a meta-analysis for that outcome?*

Author's response: Our criterion is that at least two studies must be available for a given outcome to conduct a meta-analysis for that outcome. We have included the following text in our revised manuscript: "We will perform traditional pairwise meta-analyses of all direct comparisons for which there are at least two studies available" (p.16).

4. *The investigators provide no rationale for the outcomes selected (either as primary or secondary). Interesting that no functional measures appear to be listed.*

Author's response: We would first like to clarify that based on the suggestion from reviewer 2, we have re-labelled our outcome measure 'disability' to 'physical function' (p.9). Physical function and pain intensity were selected as the primary outcomes based on recommendations

by Chiarotto et al. (2018), published in *Pain*.[2] Chiarotto et al (2018) reported the results of a Delphi study which aimed to identify the core domains that should be included in clinical trials involving non-specific low back pain. Three key domains were decided by consensus: physical functioning, pain

intensity, and health-related quality of life. We also chose to include physical function and pain intensity as our primary outcomes as they are the most responsive outcomes for measuring treatment success for chronic LBP,[3] and have also been commonly included in previous related systematic reviews.[4-7]

Health-related quality of life was selected as a secondary outcome measure it has also been identified as an important domain.[2] Fear avoidance was also selected as a secondary outcome, based on the fear avoidance model of chronic musculoskeletal pain [8-10] which describes how fear avoidance contributes to ongoing cycle of increased pain intensity and disability for chronic musculoskeletal conditions, in particular low back pain.[11] In addition, we have now also included intervention compliance as a secondary outcome measure in response to recommendations from reviewer 2, and convention for network meta-analyses to include a measure (or surrogate measure) of safety to improve the clinical utility of the results.[12]

5. *Abstract page 2, line 9. Would add "patient with" before chronic low back pain in the sentence.*

Author's response: We have added the term "people with" in the abstract (p.2).

Response to comments provided by reviewer #2

We thank reviewer 2 for the insightful comments on our manuscript. We have addressed the majority of comments provided by reviewer 2 directly in our revised manuscript. These changes can be seen in the track version of our revised manuscript. Given the volume of feedback provided, and in accordance with our correspondence with the Assistant Editor, we have focussed on the key areas of concern and provided point-by-point replies below.

INTRODUCTION

Author's response: We have restructured our introduction to rearrange the flow of the information. We have included descriptions of psychological interventions, in the context of their use in low back pain. We have also included a description of a network meta-analysis and have mentioned the assumptions underpinning valid interpretation of network meta-analysis results as suggested. We have also removed statements suggesting that clinical uncertainty in research creates a lack of confidence in clinicians (p.4-7).

METHODS AND ANALYSIS

1. *Pg 6, Line 50: Your statement that you will include '… any comparison interventions' is misleading, in my view, because you later define the set of interventions. '… any comparison interventions' implies you will include all interventions. Please clarify?*

Author's response: We intend to include any comparison interventions (e.g. all interventions). We have removed the definitions previously provided for the comparison interventions. The revised manuscript has been corrected with the following text: "There will be no restriction on the non-psychological co-interventions or comparison interventions identified by our search strategy" (p.9). However, we have now included an expanded description of how we plan to classify the comparison interventions (p.14).

2. *(i) Pg 6, Line 56: Why are you not considering unpublished data sources?*

   (ii) Pg 6, Line 56: Do you mean that you will include only those records for which you are able to retrieve a full-text? Or that you will only consider published articles that are full-length journal articles? What about short reports, research letters and conference abstracts?

   (iii) Pg 8, Line 34-41: Please clarify your intent here? You are not including unpublished reports of randomised trials? So it would appear you are searching registries to identify additional data? If so, how will you link the multiple records for each trial and prioritise the use of data from these records? What will you do when data sources conflict? If you are only searching registries to identify registered trials that have been published, why bother? This is unlikely to retrieve published trials that were not retrieved by database searches because the linkage of published records to trial registrations is imperfect. And why not include unpublished records, given that you will screen them?

Author's response: Although grey literature would provide us with a broader understanding of the available evidence, we have chosen to exclude unpublished data as data published in peer-reviewed journals typically undergo greater scrutiny of the methodology and findings through the peer review process.. By only including full-length articles published in peer-reviewed journals, this will minimise the risk of including low quality evidence in our review. Full-length articles will also provide us with the capacity to scrutinise the intervention descriptions for accurate treatment node classifications and research methodologies for risk of bias assessments. We are confident that our search strategy is comprehensive, as it expands on the search strategies and terms used in previous related systematic reviews. We have also planned to search reference lists and perform citation tracking of included studies and relevant systematic reviews to minimise omission. Therefore, we agree with reviewer 2 that is not necessary to proceed with searching registries to identify registered trials. We have revised our manuscript as follows: "Observational studies, non-randomised trials, short reports, research letters, conferences abstracts and studies that have not been published as full-length articles in peer-reviewed scientific journals will be excluded" (p.7).

3. *Pg 7, Line 18-19: Where in Neilson & Weir (2001) is this definition? I can't find it. I*

am concerned that this definition is insufficient for your purpose here. Do you mean to say that any clinical interaction that influences the psychological experience is an intervention of interest for this review? By that reasoning, what interventions do we not include? What will you do with interventions identified during the review that meet these criteria, but are not listed here? I strongly encourage you to clarify this definition to provide reasonable: 1. justification for the list of interventions as it stands 2. guidance for including interventions identified during the review that are not listed here.

Author's response: We have now provided a more detailed definition for psychological interventions in our revised manuscript and provided justification for the list of interventions included. We have also provided guidance for including other psychological interventions identified by our search strategy but not listed in the protocol or Appendix. Further, we have also updated the list of interventions to better reflect our interventions of interest and search strategy. All changes described can been seen in the sub-heading *Types of Interventions,* in the eligibility criteria section of our revised manuscript (p.8-9).

4. *The outcomes are missing the definition of the construct that is being measured. Consequentially, it is not clear what, for example, form of disability is being referred to. I infer from the outcome measures that this is low back-specific disability. I encourage the authors to re-phrase this as back-specific function and to define each outcome using: domain, construct, measure(s), time point(s). Disability might thus be written as 'Function, defined as low back-specific function, measured using the ODI, …., at xxxxx (time point).*

Author's response: We have now provided a clearer definition of the outcome measures. We thank reviewer 2 for the suggestion to change "disability" to "function." An example from our revised manuscript: "Physical function, defined as lower back specific physical function, measured at the end of treatment. Physical function is commonly measured by continuous, self-report scales (e.g. Oswestry Disability Index (ODI), Roland Morris Disability Questionnaire (RMDQ), Core Outcome Measures Index (COMI), Quebec Back Pain Disability Index (QBPDI)) or rating scales within a composite measure (e.g. 12-Item or 36-Item Short Form (SF-12, SF-36)). We will not exclude studies that use other measurement tools" (p.9).

5. *Please justify your choice of the time points for analysis? It is important that these are set at a time point at which it is reasonable for change in outcome under each intervention to have occurred. I am not experienced with all of these interventions. Do you think it is reasonable for any hypothetical participant to have experienced a change in outcome under any one of these interventions by each of the time points that you specify?*

Author's response: We chose these timepoints for analysis based on the previous Cochrane review of behavioural treatment for chronic low back pain conducted by Henschke et al (2010). Furthermore, to date, we have extracted data from approximately 30 studies and have found that these time points reasonably capture the typical treatment duration of typical psychological interventions, and the follow-up time points in these studies. We have renamed our timepoints in our revised manuscript: "Data will be classified and assessed at the following time-points:

(1) pre-intervention; (2) post-intervention (i.e. timepoint closest to end of treatment); (3) short-term treatment sustainability (≥2 months but <6 months post-intervention); (4) mid-term treatment sustainability (≥6 months but <12 months post-intervention); (5) long-term treatment sustainability (≥12 months post-intervention)…" (p.12).

6. *Lastly, I am interested in your rationale for not evaluating safety in some manner? Even acceptability of therapy (perhaps defined as all cause discontinuation during the intervention period) would be helpful?*

Author's response: We thank reviewer 2 for bringing this to our attention. We have now included intervention compliance as a secondary outcome in our revised manuscript. The following text has been included in the revised manuscript describing *types of outcome measures*: "Intervention compliance, measured as the proportion of participants randomised to the intervention group who completed the intervention during the intervention period" (p.9). In the *data extraction* section of our revised manuscript, we have included the following text: "For intervention compliance, we will extract all data on the number of participants who were randomised to the intervention group and completed the intervention. If this data is not available, we will extract the number of participants randomised to the intervention group who discontinued treatment for any reason (i.e. all-cause discontinuation) within the intervention period, to calculate the number of participants who completed treatment. We will express this data as a proportion of the total number of people randomised to the intervention group. We will extract reasons for all-cause discontinuation during the intervention period if reported. We will also extract all available data on adherence, adverse and serious adverse events, including the authors' definitions of these terms" (p.11).

Response to comments provided by reviewer #3

We would like to thank reviewer 3 for their comments.

1. *Please consider to use the PRISMA Network Meta-Analysis Extension https://protect-au.mimecast.com/s/n0BoCBNZwLiLzVNnczPJBu?domain=annals.org*

Author's response: We have used the recommended PRISMA Network Meta-Analysis Extension in the revised manuscript. We have provided a PRISMA-NMA checklist with our revised manuscript submission.

2. *Page 6, line 38 to 45: please be more specific for the proposed hierarchy, how will you decide on the order if none of the first four disability measures are present (and more than one is presented)?*

Author's response: We have included a more comprehensive description of how we will decide the hierarchy for inclusion. As an example, the following text has been included in our revised manuscript: "For the following outcomes, we will extract all available data in the order which the measurement tools are listed, in accordance with the proposed hierarchy for analysis. If a given outcome is measured by several measurement tools not explicitly listed, the hierarchy for analysis will be decided by consensus from the reviewers. For studies measuring physical function: ODI; RMDQ; COMI; QBPQI; rating scale for disability from a composite measure of physical function (e.g. SF-12, SF-36); other measurement tools" (p.11-12).

3. *Page 6, line 47 to 50: please be more specific for the proposed hierarchy: what will you choose if no NRS but several pain rating scales are present? What will you do if different time-points (pain last 3 months, last week, etc.) are present. What will you do if different pain intensity scales are present regarding average pain, worst pain etc.*

Author's response: We have included a more comprehensive description of how we will decide the hierarchy for inclusion for pain intensity. In addition to the text included in our response for question 2, "For studies measuring pain intensity: NRS; 100mm VAS; 10cm VAS; rating scale for pain intensity from a composite measure of pain intensity; other measurement tools. We will extract data on pain intensity at the time point closest to randomisation and end of treatment, in the order of average pain intensity (preferred); worst pain intensity, alternative measures of pain intensity. If several alternative measures of pain intensity are reported, we will calculate an average score" (p.12).

4. *Page 6 line 52 to page 7 line 5: Please be more specific for the proposed hierarchy: What do you do if only subscore (e.g. sf-36 subcomponents) are provided and no overall-score?*

Author's response: We have included the following text in our revised manuscript: "For studies measuring health-related quality of life: PROMIS-GH-10; EQ-5D; SF-36 or SF-12 (physical component summary sub-score); SF-36 or SF-12 (mental component summary sub-score); SF-36 (overall score); NHP; rating scale from a composite measure of health-related quality of life; other measurement tools. If only an overall score for the SF-36 is provided, we will contact authors for the physical and mental component summary sub-scores" (p.12). Although it commonly occurs in literature, it is not valid to use an overall score for the SF-36 as a single index of quality of life.[13] The same applies to the SF-12. Therefore, we will give preference to using the physical component summary sub-scores. If only an overall score is provided, we will contact the authors for the subscores.

5. *Page 6, line 16 to 31, (Types of interventions and comparators) and page 10, line 14 to 22: Will you consider to split interventions if inconsistency is present? See James, A., Yavchitz, A., Ravaud, P., & Boutron, I. (2018). Node-making process in network meta-analysis of nonpharmacological treatment are poorly reported. Journal of clinical epidemiology, 97, 95-102. And also : Caldwell, D. M., & Welton, N. J. (2016). Approaches for synthesising complex mental health interventions in meta-analysis. Evidence-based mental health, 19(1), 16-21.*

Author's response: We thank the reviewer for the references. We have chosen to adopt the splitting method suggested by Caldwell et al. (2016) to categorise the psychological interventions included in our review.[1] A detailed description of how we anticipate categorising the psychological interventions can be seen in the revised manuscript on page 4-6 and page 13-14. Since a splitting method will be used to categorise the treatment nodes, we do not anticipate that inconsistencies will be further dealt with be using a splitting method. We will use the procedures described in the sub-section *Assessment of Inconsistency* in our revised manuscript. If we find the splitting approach to categorise our treatment nodes is not appropriate, we will describe any post-hoc alternative network geometrics formed (i.e. by lumping) and justify the reasons to do so in the final review.

6. *How do you deal with doses of the interventions?*

Author's response: We will extract data on dosage and frequency of the interventions. Subject to availability of data, we will attempt to perform meta-regressions based on intervention parameters related to dosage, for example, by total length (in weeks) of the intervention or by total intended hours of the intervention during the intervention period (p.18).

7. *Page 7 line 7 to 14: Please be more specific, what will you do if only subscore of e.g. FABQ is provided?*

Author's response: We have included the following text in our revised manuscript: "For studies measuring fear avoidance: FABQ (physical activity scale); FABQ (work scale); FABQ (overall score); PCS, TSK; FPQ; rating scales of fear avoidance from a composite measure of fear avoidance; other measurement tools. If only an overall score for the FABQ is provided, we will contact authors for the physical activity and work sub-scores" (p.12).

8. *Page 8, line 15 to 19, please provide mutually exclusive categories: "Data will be classified according to short-term (<6 months), mid-term (6-12 months) or long-term follow-up (≥12 months)"*

Author's response: In light of comments made by reviewer 2, we have modified the categories

in our revised manuscript as follows: “Data will be classified and assessed at the following time-points: (1) pre-intervention; (2) post-intervention (i.e. timepoint closest to end of treatment); (3) short-term treatment sustainability (≥2 months but <6 months post-intervention); (4) mid-term treatment sustainability (≥6 months but <12 months post-intervention); (5) long-term treatment sustainability (≥12 months post-intervention)…” (p.12). These timepoints have been changed to mutually exclusive categories

9. *Page 8, line 55 to 59, please provide the rule on how to produce the overall judgment based on the five domain-level judgment*

Author’s response: We have included the following text in our revised manuscript: “An overall risk of bias judgement (low risk of bias, some concerns, or high risk of bias) will be made based on the five (or six) domain-level judgements, as described in Sterne et al 2019. Generally, the overall risk of bias judgement corresponds to the worst risk of bias in any of the five (or six) domains, however studies with multiple domains graded as ‘some concerns’ may be judged as high risk of overall bias” (p.15).

10. *Page 9, Data Analysis: could you please be more specific about the approach used within Stata and R (i.e. which ado or package will you use).*

Author’s response: We have included the following text in our revised manuscript: “Pair-wise meta-analysis and NMA will be performed in Stata using the metan command, and the network package and network graphs package respectively” (p.15-16). We have decided we will no longer use R.

11. *Page 9, Pairwise meta-analysis (Line 38 to 48): please consider to mention why you would not use the Hartung-Knapp-Sidik-Jonkman approach.*

Author’s response: We thank Reviewer 3 for the recommendation. We have chosen to use the Hartung-Knapp-Sidik-Jonkman approach in our revised manuscript. We have included the following text in our revised manuscript: “We will apply the khartung command to adjust for the Hartung-Knapp-Sidik-Jonkman (HKSJ) random-effects method, which has less error rates compared to the DerSimonian and Laird approach in particular across studies with greater heterogeneity and when the number of studies is small” (p.16).

12. *Page 9, summary measures: if you expect dichotomous data, please specify this in the “types of outcome measures”*

Author’s response: We anticipate that all our outcome measures will be continuous. We have

modified the text in our revised manuscript to read: “If data is reported as dichotomous, we will use odds ratios (ORs) and 95% CI” (p.16).

13. *Page 10, network meta-analysis: please consider to add the reference for frequentist*

“sucra” values Rücker, G., & Schwarzer, G. (2015). Ranking treatments in frequentist network meta-analysis works without resampling methods. BMC medical research methodology, 15(1), 58.

And please consider to indicate that your SUCRA is the p-score which is the frequentist analogy to the baesian SUCRA. (see reference Rücker & Schwarzer 2015).

Author's response: Thank you for the comment. We have added the suggested references and comments in our review (p.17).

14. *Page 10, assessment of inconsistency: please be more specific on how you implement the Bucher method (which software, which Stata ado or R package).*

Author's response: We have included the following text in our revised manuscript: "Local inconsistencies will be assessed using the Bucher method by computation of inconsistency factors and 95% CI for each triangular and quadratic loop in the network. The Bucher method will be implemented in Stata using the ifplot command" (p.17-18).

15. *Will the publication bias also be assessed within the network-meta-analysis or only in the pairwise meta-analysis?*

Author's response: We will also assess publication bias within the network meta-analysis. The following text has been included in the revised manuscript: "Publication bias in the NMA will be evaluated by visual inspection of comparison-adjusted funnel plots for asymmetry" (p.18).

REFERENCES

1. Caldwell DM and Welton NJ. Approaches for Synthesising Complex Mental Health Interventions in Meta-Analysis*. Evidence Based Mental Health*. 2016;**19**(1):16-21.

2. Chiarotto A, Deyo RA, Terwee CB, et al. Core Outcome Domains for Clinical Trials in Non-Specific Low Back Pain*. European Spine Journal*. 2015;**24**(6):1127-1142.

3. Chapman JR, Norvell DC, Hermsmeyer JT, et al. Evaluating Common Outcomes for Measuring Treatment Success for Chronic Low Back Pain*. Spine*. 2011;**36**:S54-S68.

4. Hoffman BM, Papas RK, Chatkoff DK, et al. Meta-Analysis of Psychological Interventions for Chronic Low Back Pain*. Health Psychology*. 2007;**26**(1):1.

5. Henschke N, Ostelo RW, van Tulder MW, et al. Behavioural Treatment for Chronic Low-Back Pain. *Cochrane Database of Systematic Reviews*. 2010;7(7).

6. Guzmán J, Esmail R, Karjalainen K, et al. Multidisciplinary Rehabilitation for Chronic Low Back Pain: Systematic Review. *BMJ*. 2001;**322**(7301):1511-1516.

7. van Tulder MW, Ostelo R, Vlaeyen JW, et al. Behavioral Treatment for Chronic Low Back Pain: A Systematic Review within the Framework of the Cochrane Back Review Group. *Spine*. 2000;**25**(20):2688-2699.

8. Leeuw M, Goossens MEJB, Linton SJ, et al. The Fear-Avoidance Model of Musculoskeletal Pain: Current State of Scientific Evidence. *Journal of Behavioral Medicine*. 2007;**30**(1):77-94.

9. Vlaeyen JW and Linton SJ. Fear-Avoidance Model of Chronic Musculoskeletal Pain: 12 Years On. *Pain*. 2012;**153**(6):1144-1147.

10. Vlaeyen JW and Linton SJ. Fear-Avoidance and Its Consequences in Chronic Musculoskeletal Pain: A State of the Art. *Pain*. 2000;**85**(3):317-332.

11. Chung EJ, Hur Y-G,Lee B-H. A Study of the Relationship among Fear-Avoidance Beliefs, Pain and Disability Index in Patients with Low Back Pain. *Journal of Exercise Rehabilitation*. 2013;**9**(6):532.

12. Rouse B, Chaimani A,Li T. Network Meta-Analysis: An Introduction for Clinicians. *Internal and Emergency Medicine*. 2017;**12**(1):103-111.

13. Lins L and Carvalho FM. Sf-36 Total Score as a Single Measure of Health-Related Quality of Life: Scoping Review. *SAGE Open Medicine*. 2016;**4**:2050312116671725.

VERSION 2 – REVIEW

REVIEWER	Cary Reid Weill Cornell Medicine New York, New York USA
REVIEW RETURNED	10-Mar-2020

GENERAL COMMENTS	The authors have been responsive to the concerns raised by the reviewers. 1. I agree with Reviewer 2 that inspection of the articles for safety outcomes would be reasonable. Reporting how many of the articles of your sample even report on one or more safety outcomes would be instructive for readers. My hunch would be very few bother to measure any potential adverse outcomes. 2. I am assuming that the authors will update their search to include articles published after August 2019.

REVIEWER	Matthew K Bagg Centre for Pain IMPACT, Neuroscience Research Australia; Prince

	of Wales Clinical School and New College Village, University of New South Wales, Sydney, Australia I have previously collaborated with Drs Ferreira and Dr Hayden. These collaborations were minor and prior to the work described in this manuscript.
REVIEW RETURNED	29-Apr-2020

GENERAL COMMENTS	**Introductory statement** Thank you for the opportunity to review this revised manuscript. The study described is a systematic review and planned network meta-analysis of psychological interventions for people with chronic low back pain. The manuscript is comprehensive and clearly written. This study will be an important contribution to the information available to decision-makers in this clinical area. **Review Checklist** Mark each as Yes or No or N/A Please elaborate on any 'No' answers in the free text section below. ***1. Is the research question or study objective clearly defined?*** Yes The research question is: "what is the most effective psychological intervention for chronic low back pain?". It is clearly worded. To the authors: Thank you for describing the assumptions upon which 'robust' network meta-analytic inference is based. Please make these minor changes to correct the description. "provided that the assumptions of transitivity (balanced distribution of potential effect modifiers across all *comparisons* within a network (*please cite Salanti G. Res Synth Methods. 2012;3:80–97* PubMed . // *Jansen JP, Naci H. BMC Med. 2013;11:159. // Bagg et al. J Physio 2018;64:128–132* PubMed)) and consistency (statistical agreement between *direct and indirect evidence for each comparison* (*please cite Efthimiou O, et al. Res Synth Methods. 2016;7:236–263* PubMed . // *Bagg et al. J Physio 2018;64:128–132* PubMed)) are satisfied" Please also add appropriate references throughout this paragraph. ***2. Is the abstract accurate, balanced and complete?*** Yes To the authors: Please make the following minor change: "We will conduct a random-effects NMA using a freuentist approach to estimate relative *effects for all comparisons between treatments and rank* treatments according to the mean rank and surface under the cumulative ranking curve values."

	3. Is the study design appropriate to answer the research question? Yes A SR and NMA is the appropriate design to answer this question. ***4. Are the methods described sufficiently to allow the study to be repeated?*** Yes, pending the considerations described under heading 7. To the authors: Thank you for clarifying your procedures for sampling interventions and defining nodes. I have a single further suggestion regarding the interventions. Please consider defining a subset of these interventions as a "decision set" if you are only interested in estimating the relative effects between psychological interventions. One may fit the models to networks containing all interventions (the analysis set) to maximise information, yet restrict the presentation of effect sizes (and construction of rankings) to the decision set. It depends on what you envisage decision-makers using these results for. I don't think you need to describe these sets now, as given your sampling procedures you may include additional interventions during the review. I recommend considering it for the paper describing the results. It appears from your response that you do consider leg pain an effect modifier if there is associated nerve root compromise? On that presumption, please note that the covariate set for assessment of transitivity and of consistency should be the same. Transitivity, heterogeneity and consistency are inter-related. Please use the same covariates in each of these assessments. Relatedly, I think you should inspect the distribution of 'leg pain with nerve root compromise' across network comparisons and also consider modelling it as a covariate in the NMA models. Thank you for your clear description of the hierarchies with which you will extract data. The MOS SF-36 and its variants are interesting, because they don't immediately appear to measure back-specific function. In constrast to ODI, RMDQ, QBPDS, PDI; the SF-x does not mention pain or back pain. I have encountered this measurement issue in my work and I think one can argue the case either way. Please note this as a potential source of heterogeneity? Please capitalise and change to '*Microsoft* Excel' in several places. Thank you for the clarity regarding your procedures for assessing transitivity. Please change the instances where you refer to distribution *across studies* to instead refer to distribution *across comparisons*. Transitivity does not require that study-level (or individual-level) covariate values are the same within a comparison. We often expect to see variability across studies in a comparison. If these covariates are associated with intervention effects this may cause heterogeneity. Transitivity is concerned with the distribution of this variability across comparisons. Differences in the variability of covariates (that are associated with intervention effects) across comparisons is akin to imbalance of these covariates in each group of a trial. We are, in effect, conditioning on these covariates when we examine their distributions to ensure they are similar.

Thank you for updating the description of CINeMA.
Please
cite https://doi.org/10.1371/journal.pmed.1003082 and https://doi.org/10.1002/cl2.1080
as these are the latest descriptions of CINeMA?
I'm still not comfortable with it being called a Cochrane web application, because whilst it has support from the Cochrane and Campbell collaborations, it was developed by Georgia Salanti's team at ISPM, Uni Bern, and other members of the Cochrane MIMG.

5. Are research ethics (e.g. participant consent, ethics approval) addressed appropriately?

Yes

6. Are the outcomes clearly defined?

Yes

7. If statistics are used are they appropriate and described fully?

Yes, pending the following considerations.

To the authors:
Please cite White 2015 STATA J 15(4): 951–985 (and the 2009 and 2011 versions) in reference to the network package (mvmeta). Please cite Chaimani et al. PLoS One. 2013 8(10):e76654. and Chaimani A, Salanti G 2015 STATA J 15(4): 905-950. in reference to the network graphs package.

I do not think your statement regarding the P-score is correct. Firstly, to my knowledge, SUCRA is calculated with the STATA command mvmeta::network rank. Whereas, the P-score is calculated using the R command netmeta::netrank. I am not aware that there is a command for the P-score in mvmeta. Secondly, network rank calculates the SUCRA using re-sampling (see White 2015 STATA J 15(4): 951–985). Whereas, the P-score is calculated without re-sampling.

Thank you for clarifying your assumption regarding the network heterogeneity variance parameter.
Although, I think you have conflated this with the heterogeneity variance parameters for pairwise comparisons. They are different parameters. A single parameter is estimated for the entire network, as an index of global network heterogeneity. Heterogeneity variance parameters are also estimated for each comparison, in the same way we would for a pair-wise meta-analysis. My question was about how you will do this for both the network parameter and the comparison parameters?

Thank you for the expanded description of network meta-regression.
You've specified the expected nature of the treatment.covariate interaction and direction of effect of each covariate. However, the statement of your assumptions regarding the regression co-efficients is missing. This is important, because it is an extension of transitivity to the treatment.covariate interactions. Please see doi: 10.1002/jrsm.1327 and doi: 10.1002/jrsm.1257 for elaboration.

I firmly recommend fitting side-splitting models when you fit the GDBTIM and use the

loop-specific approach. Each of the approaches to evaluate heterogeneity and consistency have limitations so I think you will have more success with this evaluation if you use them as complementary approaches.

8. Are the references up-to-date and appropriate?

Yes, pending the inclusions I have recommended elsewhere.

9. Do the results address the research question or objective?

N/A

10. Are they presented clearly?

N/A

11. Are the discussion and conclusions justified by the results

N/A

12. Are the study limitations discussed adequately?

Yes, to the extent reasonable in a protocol.

13. Is the supplementary reporting complete (e.g. trial registration; funding details; CONSORT, STROBE or PRISMA checklist)?

Yes. The authors have prospectively registered their systematic review on PROSPERO, and provided the ID number in accordance with BMJ Open policy. The authors have also provided accurately completed PRISMA, PRISMA-P and PRISMA-NMA checklists.

14. To the best of your knowledge is the paper free from concerns over publication ethics (e.g. plagiarism, redundant publication, undeclared conflicts of interest)?

Yes

15. Is the standard of written English acceptable for publication?

Yes

Statistics

1. Does this paper require specialist statistical review?
   Yes and I have performed this review

Would you be willing to review a revision of this manuscript?
Yes

Recommendation

Minor Revision

VERSION 2 – AUTHOR RESPONSE

Response to additional comments provided by reviewer #1

We would like to thank reviewer 1 for their additional comments on our revised manuscript.

2. *I agree with Reviewer 2 that inspection of the articles for safety outcomes would be reasonable. Reporting how many of the articles of your sample even report on one or more safety outcomes would be instructive for readers. My hunch would be very few bother to*

   measure any potential adverse outcomes.

Authors' response: As recommended, we have included *safety* as an additional secondary outcome

measure. In the Methods and Analysis (Eligibility criteria: Types of outcome measures) section, we have defined *safety* as*:* "…the proportion of participants who experience at least one adverse effect during the intervention period. Adverse effects will be broadly defined as any 'adverse event,' 'side effect,' 'complication,' or event resulting in discontinuation of treatment, associated with the intervention (psychological or comparison) under investigation" (p. 10, paragraph 1)

We have included a paragraph in the Methods and Analysis (Data Extraction: Results) section to describe the data that will be extracted for assessing *safety: "*For safety, we will extract all available data on adverse effects, broadly encompassing adverse and serious adverse events, side effects, complications, and all-cause discontinuation. We will extract authors' definitions and reasons for any adverse effects. We will also extract all available data, including authors' definitions, on alternative measures of safety reported in the included studies. We will extract the number of participants who experience at least one adverse effect related to the psychological or comparison intervention under investigation and express this as a proportion of the total number of participants randomised to each group respectively. We will also extract data on adherence if reported" (p. 12, paragraph 2).

3. *I am assuming that the authors will update their search to include articles published after August 2019.*

Authors' response: We will update our search prior to publication of the full paper to include articles published after August 2019. This final search date will be reported in the full paper.

Response to additional comments provided by reviewer #2

We would like to thank reviewer 2 for their additional comments on our revised manuscript.

4. *Please make these minor changes to correct the description. "provided that the assumptions of transitivity (balanced distribution of potential effect modifiers across all comparisons within a network (please cite Salanti G. Res Synth Methods. 2012;3:80–97. // Jansen JP, Naci H. BMC Med. 2013;11:159. // Bagg et al. J Physio 2018;64:128–132)) and consistency (statistical agreement between direct and indirect evidence for each comparison (please cite Efthimiou O, et al. Res Synth Methods. 2016;7:236–263. // Bagg et al. J Physio 2018;64:128–132)) are satisfied" Please also add appropriate references throughout this paragraph.*

Authors' response: We have added the suggested minor changes to the text and included the suggested references in our revised manuscript: "Integrating direct and indirect evidence increases the precision of treatment effect estimates, provided that the assumptions of transitivity (balanced distribution of potential effect modifiers across all **comparisons** within a network)**[41-43]** and consistency (statistical agreement between **direct and indirect evidence for each comparison)[43, 44]** are satisfied" (p. 7, paragraph 1).

5. *Please make the following minor change: "We will conduct a random-effects NMA using a frequentist approach to estimate relative effects for all comparisons between treatments and rank treatments according to the mean rank and surface under the cumulative ranking curve values."*

Authors' response: We have corrected the abstract to reflect the minor change suggested: "We will conduct a random-effects NMA using a frequentist approach to estimate relative **effects for all comparisons between treatments and rank** treatments according to the mean rank and surface under the cumulative ranking curve values" (p.2, paragraph 2).

6. *I have a single further suggestion regarding the interventions. Please consider defining a subset of these interventions as a "decision set" if you are only interested in estimating the relative effects between psychological interventions. One may fit the models to networks containing all interventions (the analysis set) to maximise information, yet restrict the presentation of effect sizes (and construction*

of rankings) to the decision set. It depends on what you envisage decision-makers using these results for. I don't think you need to describe these sets now, as

given your sampling procedures you may include additional interventions during the review. I recommend considering it for the paper describing the results.

Authors' response: We thank Reviewer 2 for the comment. We have included the following text in the revised manuscript: "A decision set and supplementary set will be formulated for the final review" (p. 15, paragraph 4).

2. *It appears from your response that you do consider leg pain an effect modifier if there is associated nerve root compromise? On that presumption, please note that the covariate set for assessment of transitivity and of consistency should be the same. Transitivity, heterogeneity and consistency are inter-related. Please use the same covariates in each of these assessments. Relatedly, I think you should inspect the distribution of 'leg pain with nerve root compromise'*

across network comparisons and also consider modelling it as a covariate in the NMA models. **Authors' response:** We have revised the manuscript to reflect the same set of covariates for the assessment of transitivity (p. 17, paragraph 4) and consistency (p. 18, paragraph 3). We will also inspect the distribution of sciatica (leg pain with nerve root compromise) and model it as a covariate in the NMA models (p. 19, paragraph 1).

3. *The MOS SF-36 and its variants are interesting, because they don't immediately appear to measure backspecific function. In contrast to ODI, RMDQ, QBPDS, PDI; the SF-x does not mention pain or back pain. I have encountered this measurement issue in my work and I think*

one can argue the case either way. Please note this as a potential source of heterogeneity? **Authors' response:** We thank the reviewer for bringing this to our attention. In recognition of the SF-x as a potential source of heterogeneity, we have included the following text in the relevant sections of our revised manuscript:

(iv) Under heading *Assessment of inconsistency: "…*we will examine the potential influence of the pre-specified effect modifiers within inconsistent loops using network meta-regression models or sub-group analyses, and conduct sensitivity analyses excluding studies that may be the source of inconsistency (e.g. high risk of bias, studies measuring physical function using the SF-12 or SF-36)" (p. 18, paragraph 3).

(v) Under heading *Sensitivity and sub-group analysis: "*We will also perform a sensitivity analysis by excluding studies measuring physical function using the SF-12 or SF-36, which may be a potential

source of heterogeneity, provided that sufficient data for physical function is available and the original network structure remains the same” (p. 19, paragraph 1).

6. *Please capitalise and change to 'Microsoft Excel' in several places.*

Authors' response: We have capitalised and changed to 'Microsoft Excel' in the relevant sections of the revised manuscript (p. 11, paragraph 1 & p. 15, paragraph 5).

4. *Please change the instances where you refer to distribution across studies to instead refer to distribution across comparisons. Transitivity does not require that study-level (or individual-level) covariate values are the same within a comparison. We often expect to see variability across studies in a comparison. If these covariates are associated with intervention effects this may cause heterogeneity. Transitivity is concerned with the distribution of this variability across comparisons. Differences in the variability of covariates (that are associated with intervention*

   effects) across comparisons is akin to imbalance of these covariates in each group of a trial. We are, in effect, conditioning on these covariates when we examine their distributions to ensure they are similar.

Authors' response: We have revised the relevant sections of text to read '*across comparisons'* (p.

17, paragraph 3 line 4 & p. 18, paragraph 1 lines 4 and 7).

8. *Thank you for updating the description of CINeMA. Please cite https://doi.org/10.1371/journal.pmed.1003082 and https://doi.org/10.1002/cl2.1080 as these are the latest descriptions of CINeMA? I'm still not comfortable with it being called a Cochrane web application, because whilst it has support from the Cochrane and Campbell collaborations, it was developed by Georgia Salanti's team at ISPM, Uni Bern, and other members of the Cochrane MIMG.*

Authors' response: We have incorporated the suggested citations for the latest descriptions of CINeMA (p. 20, paragraph 5 line 2). Further, we have removed the term 'Cochrane' from the description of CINeMA in the abstract and main text.

5. *Please cite White 2015 STATA J 15(4): 951–985 (and the 2009 and 2011 versions) in reference to the network package (mvmeta). Please cite Chaimani et al. PLoS One. 2013 8(10):e76654. and Chaimani A, Salanti G 2015 STATA J 15(4): 905-950. in reference to the network graphs package.*

Authors' response: We have incorporated the suggested citations in our revised manuscript: "Pair-wise meta-analysis and NMA will be performed in Stata[61] using the metan command (with

Knapp–Hartung adjustment applied), and the network package**[62-64]** and network graphs package**[65, 66]** respectively" (p. 16, paragraph 2 lines 4-5).

6. *I do not think your statement regarding the P-score is correct. Firstly, to my knowledge, SUCRA is calculated with the STATA command mvmeta::network rank. Whereas, the P-score is calculated using the R command netmeta::netrank. I am not aware that there is a command for the P-score in mvmeta. Secondly, network rank calculates the SUCRA using re-sampling (see White 2015 STATA J 15(4): 951– 985). Whereas, the P-score is calculated without re-*

   sampling.

Authors' response: We have removed the sentence regarding the p-score.

2. *Thank you for clarifying your assumption regarding the network heterogeneity variance parameter. Although, I think you have conflated this with the heterogeneity variance parameters for pairwise comparisons. They are different parameters. A single parameter is estimated for the entire network, as an index of global network heterogeneity. Heterogeneity variance parameters are also estimated for each comparison, in the same way we would for a*

   pair-wise meta-analysis. My question was about how you will do this for both the network parameter and the comparison parameters?

Authors' response: We have revised our manuscript to provide clarity, differentiating between assumptions regarding the network heterogeneity variance parameter, and the homogeneity variance parameters estimated for each pairwise comparisons.

> For pairwise comparisons: "we will assume the heterogeneity variance for each pairwise comparison is different" (p. 17, paragraph 3).
>
> For the NMA, "We will assume the heterogeneity variance across different comparisons within the NMA model will be the same.[76] We will use heterogeneity variances from the NMA model as an index of global network heterogeneity." (p. 18, paragraph 2).

3. *Thank you for the expanded description of network meta-regression. You've specified the expected nature of the treatment.covariate interaction and direction of effect of each covariate. However, the statement of your assumptions regarding the regression co-efficients is missing. This is important,*

because it is an extension of transitivity to the treatment.covariate interactions. Please see doi: 10.1002/jrsm.1327 and doi:

10.1002/jrsm.1257 for elaboration.

Authors' response: We have included the following text in our revised manuscript: "We will assume that for each network meta-regression model, the regression co-efficient for each covariate will be the same across all comparisons in the network" (p. 19, paragraph 1).

4. *I firmly recommend fitting side-splitting models when you fit the GDBTIM and use the loop-specific approach. Each of the approaches to evaluate heterogeneity and consistency have limitations so I think you will have more success with this evaluation if you use them as complementary approaches.*

Authors' response: We have revised the text in the manuscript to read as follows: "Local inconsistencies within closed loops will be assessed with the loop specific approach (Bucher method),[78] and by fitting side-splitting models.[62]" (p. 18, paragraph 3).

VERSION 3 – REVIEW

REVIEWER	Matthew K Bagg Neuroscience Research Australia and UNSW, Sydney, Australia
REVIEW RETURNED	25-Jun-2020

GENERAL COMMENTS	Thank you for the opportunity to review this revised manuscript and for engaging with my prior reviews. This protocol is clear, comprehensive and well written. The study will be an important contribution to our field. I am looking forward to reading the results.