PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Effectiveness and harms of high-flow nasal oxygen (HFNO) for acute respiratory failure: a systematic review protocol
AUTHORS	Baldomero, Arianne; Melzer, Anne; Greer, Nancy; Majeski, Brittany; Macdonald, Roderick; Wilt, Timothy

VERSION 1 – REVIEW

REVIEWER	Bram Rochwerg McMaster University, Canada
REVIEW RETURNED	22-Oct-2019

GENERAL COMMENTS	Thanks for asking me to review this systematic review protocol for BMJ Open. The review addresses an important topic, and it is clear the authors have put a considerable amount of thought into this protocol. I have a number of suggestions that I believe will improve the presentation of this work, Major Comments: 1. Other HFNO meta-analyses – a number of other MAs summarizing the efficacy and safety of HFNO (HFNC) have recently been published, including a very recent one in Intensive Care Medicine (PMID: 30888444). I think some discussion of this previous work, and how the proposed MA incrementally adds to the literature (perhaps including more recent studies, differences in eligibility or analysis, etc) would be important. 2. Study inclusion – what if potential studies include 80% adults, or hypoxia criteria that closely match the author's criteria but not exactly? Will they be excluded? I think its important to also discuss how these situations will be handled. 3. Outcomes – Given this is a protocol publication, I think being as explicit as possible with outcomes measures is crucial. a. For example, mortality at what endpoint? What if multiple endpoints for certain outcomes are reported (eg. 48 hours, 1 week, 30 days), which will the authors use for analysis? b. Are intubation rate and duration of invasive mechanical ventilation (the preferable descriptor of this outcome) separate outcomes of interest? This isn't clear as currently written. c. How is the outcome of ICU transfer different from ICU admission? d. To meet inclusion does delirium have to be diagnosed using a validated tool or is this completely dependent on individual study definitions (ie. If they report delirium it will be included no matter what). e. Will all measures of barotrauma be combined together into a single outcome or pooled separately in analysis? Same question for gastric dysfunction and functional independence at discharge. f. Do you have specific criteria for compromised nutrition that would meet eligibility for this outcome?

	g. I'm certain that multiple endpoints for the intermediate outcomes will be reported, the authors should clarify at what endpoint they will pool data. What about studies report only PaO2 and not SpO2? 4. Study Design – will the authors include crossover RCTs? If yes, how will these be analyzed? What about cluster RCTs if they are found? 5. Subgroup analysis – this section of the methods requires revision for clarity. Statistical heterogeneity, if found, should be explored based on a priori subgroup analyses. The methods section requires a distinct section dedicated to subgroup analysis. The authors should prioritize which subgroup analysis they will perform, if sufficient data allows, and explicitly list these as clearly as possible in terms of how they will identify the subgroups of interest. Examples may include: 1) COT vs NIV as comparator, 2) hypoxic vs hypercarbic respiratory failure (with clear criteria for both)….etc. etc. These are only examples, the authors may prioritize others. Also, each proposed subgroup analysis should include an a priori hypothesis of effect, eg. we hypothesize that HFNO will be more beneficial in hypoxic respiratory failure as opposed to hypercarbic. A priori hypothesis limit the potential for spurious subgroup findings. Finally, the authors should describe how they will assess for credible subgroup effects (eg. Chi-squared test and what p-value threshold they will use). 6. Sensitivity analysis – also a tool to explore unanticipated heterogeneity, involves excluding studies based on a certain eligibility decision to see whether it changes the overall results. The methods section should clearly state if any a priori sensitivity analyses are planned. A common consideration would be sensitivity analyses excluding studies that were only reported in abstract form. Although some systematic review authors account for risk of bias by performing sensitivity analyses excluding studies judged to be at high risk of bias, my personal preference is to present subgroup analysis based on low vs high risk of bias studies (hypothesizing that effect will be greater in high risk of bias studies). That being said, and my personal preference aside, either approach is reasonable. The description of sensitivity analysis in the methods requires revision for clarity. Upon further review, this is described properly in the 'assessment of bias in individual studies' section, but not described properly in the 'data synthesis and analysis' section. Minor Comments: 1. Tense – the verb tense is variable throughout the manuscript (eg. "we searched…". Whether the authors choose the future or past tense (future may be preferable given this is a protocol), at the very least it should be consistent. 2. Abstract, methods – given this is a protocol, more space should be dedicated to analytic methods, risk of bias assessment (if it is planned), certainty in pooled effect estimate assessment (if it is planned), etc. 3. Strengths & Limitations – typo, replace 'effectives' with 'efficacy' 4. Strengths & Limitations – I'm not sure I understand the last point regarding subgroups, isn't the goal of subgroup analyses to explore this relative heterogeneity in PICO components? Why would this preclude meaningful analyses? 5. PROSPERO – please include registration number when known 6. Methods, Timing – I presume the authors will also include patients admitted for other reasons than ARF if they develop ARF during the hospitalization? 7. Search Dates – the authors have decided to search from 2000 until present, although I agree with the decision, it would be optimal

	to include rationale for this decision in the protocol (HFNO not widely used in adults prior to 2000) 8. Conference Proceedings – it is important to be explicit regarding exactly which conference proceedings and which years will be searched. 9. Two Data abstractors – the authors should consider duplicate and independent data extraction rather than extraction and verification as the former has been demonstrated to be much more proficient at identifying extraction errors 10. Methods, Data extraction and management – I'm not certain what 'stratified data extraction' means. I presume data will be extracted similarly for all eligible studies, and then subgroup analyses will be performed. If this is correct, I would explain it like this. 11. Assessment of heterogeneity – optimal methods for examining heterogeneity according to Cochrane methods include the Isquared statistic (not test), the Chi-squared test and visual inspection of the forest plots, I'd suggest to revise to make this clear. 12. Risk of bias tool – the authors state a modified RoB tool will be used however the traditional risk of bias tool is referenced. Moderate risk of bias is described in the analysis section however is not one of the outputs of the traditional RoB tool. 13. Publication bias assessment – although not anticipated, if sufficient studies are found, will the authors then perform funnel plot analysis for publication bias?

REVIEWER	Andrea Cortegiani University of Palermo, Italy
REVIEW RETURNED	24-Oct-2019

GENERAL COMMENTS	ABSTRACT: 1) Line 9-12 I disagree on the fact that evidence on the use of HFNT in ARF in adults hospitalized patients is lacking (e.g. doi 10.1007/s00134-019-05590-5 OR doi: 10.1016/j.jcrc.2018.12.015). I can suggest changing this sentence in "the evidence is debated" or "unclear". 2) Line 30-32 From this sentence I understand that the comparator is this COT or NIV (whatever patients received). Several meta-analysis did like this in various setting and there is a huge discussion on the appropriateness of this comparator. COT and NIV are completely different interventions and so would be presented separately (HFNO vs NIV and HFNO vs COT). If this is the plan of the authors, this should be specified also in the abstract. 3) Which mortality? At which timepoint? 4) The authors should better specify what they mean when using the terms "hospitalized patients". Even patients in ICU are hospitalized. So they will include any patients admitted in hospital, whatever the department? INTRODUCTION 1) Line 14-17 I can suggest supporting this sentence about the characteristics and effect of HFNO with references (e.g. 10.1016/j.tacc.2019.02.001 ; doi: 10.1186/s12871-018-0623-4) 2) About the comfort, the evidence is unclear (doi:10.1186/s13054-019-2473-y) 3) Please specify that the PEEP effect if of low (very low) amount and depend on mouth closing METHODS Line 51-56 pag 8 could you specify if you are going to creat one single comparator (NIV and COT) or separate analysis from the

	beginning? OUTCOME MEASURE 1) It is important to state which mortality you are going to evaluate TIMING 1) Stating like this, the authors are excluding also HFNT used as "prophylaxis" of reintubation. They will lose many studies. Please clarify

VERSION 1 – AUTHOR RESPONSE

Reviewer: 1
Reviewer Name: Bram Rochwerg
Institution and Country: McMaster University, Canada
Please state any competing interests or state 'None declared': None declared.

Thanks for asking me to review this systematic review protocol for BMJ Open. The review addresses an important topic, and it is clear the authors have put a considerable amount of thought into this protocol. I have a number of suggestions that I believe will improve the presentation of this work.
Thank you very much for your very thorough review of our protocol. Your comments and suggestions are very helpful and will help us clarify our methods.

Major Comments:
1. Other HFNO meta-analyses – a number of other MAs summarizing the efficacy and safety of HFNO (HFNC) have recently been published, including a very recent one in Intensive Care Medicine (PMID: 30888444). I think some discussion of this previous work, and how the proposed MA incrementally adds to the literature (perhaps including more recent studies, differences in eligibility or analysis, etc) would be important.
We updated the introduction by providing a stronger rationale for our systematic review. Compared to the existing reviews in this area, our systematic review will compare high-flow nasal oxygen (HFNO) to both noninvasive ventilation and conventional oxygen therapy, assess a wider range of clinical conditions (e.g. chronic obstructive pulmonary disease, cardiogenic pulmonary edema, immunosuppressed, post-surgery, post-extubation, etc.) in multiple clinical settings (emergency department, intensive care unit, intermediate/step-down unit, and hospital ward), and evaluate a more comprehensive list of key clinical outcomes. In addition, there has been an increase in the number of HFNO clinical trial publications in the last couple of years warranting the need for an updated review. Our systematic evidence review will be used by the American College of Physicians-Clinical Guidelines Committee in developing a clinical practice guideline for the use of HFNO in managing acute respiratory failure.

2. Study inclusion – what if potential studies include 80% adults, or hypoxia criteria that closely match the author's criteria but not exactly? Will they be excluded? I think its important to also discuss how these situations will be handled.
We will include studies if at least 75% of the participants meet the inclusion criteria. Regarding respiratory failure criteria, we will use the mean (or median) values of the reported physiologic parameters (SpO2, PaO2, PaO2:FiO2 ratio, or PaCO2) as we have pre-specified in our inclusion criteria. We will include the study if at least one criterion for respiratory failure is met. We updated the Eligibility criteria section to reflect this.

3. Outcomes – Given this is a protocol publication, I think being as explicit as possible with outcomes measures is crucial.
We have updated our Outcomes measures section to be increase outcome measure specificity.

a. For example, mortality at what endpoint? What if multiple endpoints for certain outcomes are reported (eg. 48 hours, 1 week, 30 days), which will the authors use for analysis?
If multiple time points are reported for mortality, we will report “in-hospital” and the longest available through 90 days. If multiple points are reported for other outcomes such as patient comfort and physiologic (intermediate) outcomes, we will categorize these as “short” (first time point) and “longer” (last time point) term outcomes.

b. Are intubation rate and duration of invasive mechanical ventilation (the preferable descriptor of this outcome) separate outcomes of interest? This isn’t clear as currently written.
We will report intubation rate and duration of invasive mechanical ventilation as separate outcomes.

c. How is the outcome of ICU transfer different from ICU admission?
We will report ICU admission (for example, a patient from the ED admitted to the ICU) and ICU transfer (for example, hospital ward or step down patient transferred to the ICU for escalation of care) as ICU admission and/or transfer and if available we will also report mean number of days for ICU length of stay.

d. To meet inclusion does delirium have to be diagnosed using a validated tool or is this completely dependent on individual study definitions (ie. If they report delirium it will be included no matter what).
We will report delirium diagnosed using a validated tool and according to study definitions. We will specify the method used for delirium inclusion in our data extraction.

e. Will all measures of barotrauma be combined together into a single outcome or pooled separately in analysis? Same question for gastric dysfunction and functional independence at discharge.
We will report and analyze measures of barotrauma separately. We will analyze gastric dysfunction as a single outcome that will include placement of a nasogastric tube for decompression or treatment of abdominal distension, nausea, or vomiting; we will describe the definitions used by authors. If studies report different measures of functional independence, we will use standardized mean differences to enable pooling, if appropriate, and/or comparisons across studies.

f. Do you have specific criteria for compromised nutrition that would meet eligibility for this outcome?
We will use ‘enteral or parenteral nutrition’ as the specific criteria for compromised nutrition.

g. I’m certain that multiple endpoints for the intermediate outcomes will be reported, the authors should clarify at what endpoint they will pool data. What about studies report only PaO2 and not SpO2?
If multiple points are reported for other outcomes such as patient comfort and physiologic (intermediate) outcomes, we will report the “short” term (first time point) and the “longest” term (last time point). We will also explore analyses based on commonly reported time points. We will extract and report all intermediate/physiologic outcomes (PaO2, SpO2, PaO2:FiO2 ratio, pH, PaCO2) reported by the study.

4. Study Design – will the authors include crossover RCTs? If yes, how will these be analyzed? What about cluster RCTs if they are found?
We will include crossover RCTs, but only include patient comfort/dyspnea and intermediate outcomes from those studies. We will include cluster RCTs if they are found and we will report outcomes separately from non-cluster RCTS. We will evaluate cluster RCTs for statistical measures that adjust for clustering. We will consider sensitivity analyses based on study design.

5. Subgroup analysis – this section of the methods requires revision for clarity. Statistical heterogeneity, if found, should be explored based on a priori subgroup analyses. The methods section requires a distinct section dedicated to subgroup analysis. The authors should prioritize which

subgroup analysis they will perform, if sufficient data allows, and explicitly list these as clearly as possible in terms of how they will identify the subgroups of interest. Examples may include: 1) COT vs NIV as comparator, 2) hypoxic vs hypercarbic respiratory failure (with clear criteria for both)….etc. etc. These are only examples, the authors may prioritize others. Also, each proposed subgroup analysis should include an a priori hypothesis of effect, eg. we hypothesize that HFNO will be more beneficial in hypoxic respiratory failure as opposed to hypercarbic. A priori hypothesis limit the potential for spurious subgroup findings. Finally, the authors should describe how they will assess for credible subgroup effects (eg. Chi-squared test and what p-value threshold they will use).
If sufficient data allows, we plan to perform analysis on the following subgroups of interest: (1) noninvasive ventilator (NIV) vs. conventional oxygen therapy (COT); (2) ED, ICU, hospital ward/step down, or mixed settings; (3) chronic obstructive pulmonary disease (COPD), cardiogenic pulmonary edema/acute decompensated heart failure, pneumonia, obese, post-extubation, post-surgical, immunocompromised, (4) hypoxic, hypercapnic, and mixed (hypoxic or hypercapnic) respiratory failure; (5) treatment duration <6 vs. ≥6 hours; and (6) lower (≤30 L/min) vs. higher (>30 L/min) flow settings.

We hypothesize that: (1) HFNO is more beneficial than COT, but is as effective, though less comfortable, than NIV; (2) the efficacy of HFNO is likely the same as NIV, but better than COT, in different settings; (3) HFNO is as effective as NIV in COPD, pneumonia, post-extubation, and post-surgical patients; (4) HFNO is less effective than NIV in cardiogenic pulmonary edema and obesity due to lower level of PEEP; (5) HFNO is more effective than COT in most disease states; (6) HFNO is more effective and less harmful than NIV in hypoxic respiratory failure, but is less effective in hypercapnic and mixed hypoxic and hypercapnic respiratory failure; and (7) higher flow (>30 L/min) is more effective, but is less comfortable, than lower flow (≤30 L/min) settings. If subgroup analyses are performed, we will assess subgroup effects with an I-squared statistic for subgroup differences. The I-squared statistic delineates the percentage of variability in the estimates of effect between the different subgroups that is due to real subgroup differences (as opposed to sampling error).

6. Sensitivity analysis – also a tool to explore unanticipated heterogeneity, involves excluding studies based on a certain eligibility decision to see whether it changes the overall results. The methods section should clearly state if any a priori sensitivity analyses are planned. A common consideration would be sensitivity analyses excluding studies that were only reported in abstract form. Although some systematic review authors account for risk of bias by performing sensitivity analyses excluding studies judged to be at high risk of bias, my personal preference is to present subgroup analysis based on low vs high risk of bias studies (hypothesizing that effect will be greater in high risk of bias studies). That being said, and my personal preference aside, either approach is reasonable. The description of sensitivity analysis in the methods requires revision for clarity. Upon further review, this is described properly in the ‘assessment of bias in individual studies’ section, but not described properly in the ‘data synthesis and analysis’ section.
A description of sensitivity analyses is described in lines 262-268 under the ‘data synthesis and analysis’ section.

Minor Comments:
1. Tense – the verb tense is variable throughout the manuscript (eg. “we searched…”. Whether the authors choose the future or past tense (future may be preferable given this is a protocol), at the very least it should be consistent.
We chose future tense throughout the manuscript.

2. Abstract, methods – given this is a protocol, more space should be dedicated to analytic methods, risk of bias assessment (if it is planned), certainty in pooled effect estimate assessment (if it is planned), etc.
We updated the abstract to provide more details to describe the analytic methods. We attempted to

provide more details to describe risk of bias assessment and certainty in pooled effect estimate assessment, however, we were limited by the abstract word count.

3. Strengths & Limitations – typo, replace ‘effectives’ with ‘efficacy’
4. Strengths & Limitations – I’m not sure I understand the last point regarding subgroups, isn’t the goal of subgroup analyses to explore this relative heterogeneity in PICO components? Why would this preclude meaningful analyses?
We revised our strengths and limitations as suggested by the Editorial team.

5. PROSPERO – please include registration number when known
We updated the PROSPERO registration number.

6. Methods, Timing – I presume the authors will also include patients admitted for other reasons than ARF if they develop ARF during the hospitalization?
Yes, we will include “patients hospitalized for ARF or who developed ARF while hospitalized” as described in the Methods section, Timing subsection.

7. Search Dates – the authors have decided to search from 2000 until present, although I agree with the decision, it would be optimal to include rationale for this decision in the protocol (HFNO not widely used in adults prior to 2000)
We added the rationale for our decision to search from 2000 until present in the data sources and search strategy section: “We will search MEDLINE®, Embase, CINAHL, and Cochrane Library from January 2000 to August 2019. HFNO was not widely used in adults prior to 2000.”

8. Conference Proceedings – it is important to be explicit regarding exactly which conference proceedings and which years will be searched.
We revised the data sources and search strategy section to only include searching ClinicalTrials.gov for publications from recently completed or on-going trials. We will not be including conference proceedings.

9. Two Data abstractors – the authors should consider duplicate and independent data extraction rather than extraction and verification as the former has been demonstrated to be much more proficient at identifying extraction errors
We considered this. We routinely conduct single reviewer data abstraction with independent second data review. We conduct extensive data extractor training, develop and pilot test data extraction forms and discuss clinical and methodological concerns. This methodology is widely used by the Agency for Healthcare Research and Quality (AHRQ)-Evidence-based Practice Centers and other nationally funded systematic review groups. All results are reviewed by a third (senior author) investigator and in group meetings.

10. Methods, Data extraction and management – I’m not certain what ‘stratified data extraction’ means. I presume data will be extracted similarly for all eligible studies, and then subgroup analyses will be performed. If this is correct, I would explain it like this.
We updated the data extraction and management section to clarify this: “Data will be extracted similarly for all eligible studies and then subgroup analyses will be performed.”

11. Assessment of heterogeneity – optimal methods for examining heterogeneity according to Cochrane methods include the Isquared statistic (not test), the Chi-squared test and visual inspection of the forest plots, I’d suggest to revise to make this clear.
We included this. We regularly assess for statistical heterogeneity by examining and reporting the I-squared test, the Chi-squared test and visual inspection of forest plots. We also limit statistical pooling to studies and outcomes where there is sufficient information and they are judged to be clinically

homogeneous. We conduct sensitivity analyses to explore findings when statistical heterogeneity exists.

12. Risk of bias tool – the authors state a modified RoB tool will be used however the traditional risk of bias tool is referenced. Moderate risk of bias is described in the analysis section however is not one of the outputs of the traditional RoB tool.
Our risk of bias assessment is based on the referenced tool with individual elements rated low, unclear, or high risk of bias. Our modification of the tool is to identify overall study risk of bias as low, moderate, or high. A study with unclear elements will be considered moderate risk of bias.

13. Publication bias assessment – although not anticipated, if sufficient studies are found, will the authors then perform funnel plot analysis for publication bias?
Yes, we will perform funnel plot analysis for publication bias across studies if sufficient studies are found. We updated the Methods section to clarify this.

Reviewer: 2
Reviewer Name: Andrea Cortegiani
Institution and Country: University of Palermo, Italy
Please state any competing interests or state 'None declared': None declared

ABSTRACT: 1) Line 9-12 I disagree on the fact that evidence on the use of HFNT in ARF in adults hospitalized patients is lacking (e.g. doi 10.1007/s00134-019-05590-5 OR doi: 10.1016/j.jcrc.2018.12.015). I can suggest changing this sentence in "the evidence is debated" or "unclear".
We updated the abstract introduction and the introduction sections.

2) Line 30-32 From this sentence I understand that the comparator is this COT or NIV (whatever patients received). Several meta-analysis did like this in various setting and there is a huge discussion on the appropriateness of this comparator. COT and NIV are completely different interventions and so would be presented separately (HFNO vs NIV and HFNO vs COT). If this is the plan of the authors, this should be specified also in the abstract.
We will perform separate analyses to compare HFNO vs. NIV and HFNO vs. COT. We updated the abstract and the methods sections to clarify this.

3) Which mortality? At which timepoint?
If multiple time points are reported for "all-cause" mortality, we will report in-hospital and the longest available within 90 days.

4) The authors should better specify what they mean when using the terms "hospitalized patients". Even patients in ICU are hospitalized. So they will include any patients admitted in hospital, whatever the department?
We will include studies that randomized patients in the hospital (including hospital wards, intermediate/step-down units, and ICUs) and emergency departments. We describe this in our Methods section, Setting subsection. We are not able to add this detail to our abstract due to word count limitation.

INTRODUCTION
1) Line 14-17 I can suggest supporting this sentence about the characteristics and effect of HFNO with references (e.g. 10.1016/j.tacc.2019.02.001 ; doi: 10.1186/s12871-018-0623-4)
We added these references.

2) About the comfort, the evidence is unclear (doi:10.1186/s13054-019-2473-y)

We added this reference.

3) Please specify that the PEEP effect if of low (very low) amount and depend on mouth closing
We revised the Introduction to reflect this low PEEP effect.

METHODS
Line 51-56 pag 8 could you specify if you are going to creat one single comparator (NIV and COT) or separate analysis from the beginning?
We specified that we will be performing separate analysis for HFNO vs. NIV and HFNO vs. COT in the Eligibility criteria section, Comparators subsection.

OUTCOME MEASURE
1) It is important to state which mortality you are going to evaluate
We will report all-cause mortality. If multiple time points are reported for mortality, we will report “in-hospital” and the longest available through 90 days.

TIMING
1) Stating like this, the authors are excluding also HFNT used as "prophylaxis" of reintubation. They will lose many studies. Please clarify.
Based in part on discussion of our end-users (American College of Physicians-Clinical Guideline Committee), the purpose and scope of this review is the management of patients with respiratory failure rather than the prophylactic role of HFNO on reintubation rates. We will only be including studies that meet criteria for respiratory failure. Studies that included patients with respiratory failure post-extubation will be included. We agree that an additional review that addresses the role of prophylaxis would be useful.

VERSION 2 – REVIEW

REVIEWER	Bram Rochwerg McMaster University, Canada
REVIEW RETURNED	28-Nov-2019

GENERAL COMMENTS	Thanks to the authors for addressing my comments and those of the other peer reviewer in a careful and thoughtful manner. I am certain the manuscript is now stronger and the methods more explicitly defined. I have a few residual comments, 1) Subgroup analysis - this section is much improved. However, issues remain. Cochrane guidance for assessing subgroup differences is to use the Chi-squared test (not the I-squared statistic) however this is only useful if there are 2 subgroups. For some of the subgroups, the authors have listed multiple potential subgroups, and as far as I'm aware the only method for comparing multiple subgroups in the same analysis is meta-regression. 2) Data synthesis and analysis - again, although improved greatly, inaccuracies in this section remain. If heterogeneity is high, the next step is most commonly subgroup analysis, and not sensitivity analysis. If you choose to do sensitivity analysis based on risk of bias, you will exclude high risk of bias studies in the sensitivity analysis not 'include them'.

REVIEWER	Andrea Cortegiani University of Palermo, Italy
REVIEW RETURNED	30-Nov-2019

GENERAL COMMENTS	I have no further queries.

VERSION 2 – AUTHOR RESPONSE

1) Subgroup analysis - this section is much improved. However, issues remain. Cochrane guidance for assessing subgroup differences is to use the Chi-squared test (not the I-squared statistic) however this is only useful if there are 2 subgroups. For some of the subgroups, the authors have listed multiple potential subgroups, and as far as I'm aware the only method for comparing multiple subgroups in the same analysis is meta-regression.

We updated the data synthesis and analysis section. For analyses involving two subgroups, the chi-squared test will be useful to assess differences between the groups. If applicable, when there are more than two subgroups, meta-regression will be applied to explore the relationship between the subgroup characteristics and the treatment effects [1]. Meta-regression will only be considered if there are more than ten studies in a meta-analysis. Meta-regression will be performed using the 'metafor' package for R.

Reference

1. Thompson SG, Higgins JP: How should meta-regression analyses be undertaken and interpreted? *Stat Med* 2002, 21(11):1559-1573.

2) Data synthesis and analysis - again, although improved greatly, inaccuracies in this section remain. If heterogeneity is high, the next step is most commonly subgroup analysis, and not sensitivity analysis. If you choose to do sensitivity analysis based on risk of bias, you will exclude high risk of bias studies in the sensitivity analysis not 'include them'.

We updated the data synthesis and analysis section: "If heterogeneity exists, we will conduct subgroup analyses to explore potential causes of heterogeneity."

Regarding sensitivity analysis based on risk of bias, we respect the input of the reviewer. The preferred approach for presenting data that includes high risk of bias is not certain and varies. Based on prior methodological guidance, we plan to present our base findings that include studies deemed low or moderate risk of bias as that information is most likely to be credible. We will conduct and report sensitivity analyses that includes data from studies deemed to be high risk of bias. We updated the data synthesis and analysis section to clarify this: "Our primary analysis will include studies deemed of low to moderate risk of bias. We will conduct sensitivity analyses that includes data from studies deemed to be high risk of bias."