PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form (**http://bmjopen.bmj.com/site/about/resources/checklist.pdf**) and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

ARTICLE DETAILS

TITLE (PROVISIONAL)	PATIENT SAFETY DURING PROCEDURAL SEDATION USING CAPNOGRAPHY MONITORING: A SYSTEMATIC REVIEW AND META-ANALYSIS
AUTHORS	Saunders, Rhodri; Struys, Michel; Pollock, Richard; Mestek, Michael; Lightdale, Jenifer

VERSION 1 - REVIEW

REVIEWER	Melissa Langhan Associate Professor Yale University New Haven, CT, USA
REVIEW RETURNED	19-Jul-2016

GENERAL COMMENTS	This was a very well written manuscript. My only comment would be to consider clarifying the primary endpoints - they are not well defined in the methods, and with a small number of eligible studies, it may be feasible to contact the authors for more information on specific endpoints to improve homogeneity of results.

REVIEWER	Kim van Loon University Medical Center Utrecht, The Netherlands PhD student and specialist registrar in Anaesthesia
REVIEW RETURNED	26-Jul-2016

GENERAL COMMENTS	The authors of this manuscript performed a systematic review and meta-analysis to answer the research question whether adding capnography to pulse oximetry and visual assessment would reduce sedation related events during procedural sedation. First I would like to congratulate the authors with the manuscript, it is well written and gives a comprehensive overview of RCT's in this field. However, I also have a few major- and some minor comments that I would like to share with the authors. 1. On page 6, the authors state that they use the modified Jadad score to assess the study quality. The score was developed to evaluate quality of clinical reports in pain relief, thus for intervention studies evaluating treatment. In the Jadad score, double blinding is an important quality indicator. However, intervention studies in this setting compare two monitoring strategies, not therapy alone. Meaning they do not only assess the additional monitor, but also the alarm strategy, how the caregivers respond and what they do (therapy) with the additionally received information. In real life, capnography readings are interpreted by the caregivers and not by an independent observer in a semi-experimental setting. The

	interpretation of multiple monitors, that sometimes give contradictory information, can be demanding and confusing. Therefore, I consider the quality indicators chosen not fully applicable to intervention studies evaluating monitoring strategies. This is consistent with the reasoning of Conway et al. (p12 line 25-45) I would suggest modifying the quality score or alternatively describing this more clearly as a limitation in the discussion. 2. In the sensitivity analysis the authors add 7 scenarios after they have evaluated several outcomes (e.g. mild and severe hypoxemia, apnoea) with subgroups of endpoint definitions and study qualities. That adds up to a lot of comparisons that could potentially result in a type I error. As the authors performed mixed effect models, they could consider, when they consider these scenarios (p7 line 23-30) as important confounders, to add them as fixed effects in the mixed models. 3. I would consider “the administration of standard vs. rescue supplemental oxygen” as an important confounder for the determinant-outcome relation (the research question) as stated on p4 line 48-55. I would suggest adding this characteristic in table 1, and if possible also as a fixed effect in the mixed models (I could imagine that the model is not able to converge with too many fixed effects). Furthermore, the characteristic “who was responsible for monitoring the patient (e.g dedicated nurse, doctor)” would be a nice complement to table 1. Minor comments: - The abstract, p2 line 10, PSA is not fully written in the abstract. - p4, line 11, PSA definition could also be considered for surgical procedures, not provided only for non-surgical procedures. - p5, line 24, you could consider using Web of Knowledge to assure that you are not missing any important papers by screening the references and citations of the included papers. - p12, line 57, I would refrain from using reference 26. The paper of Ehrenfeld et al. discussed hypoxemia in the surgical population, a completely different setting and population from the PSA population. Furthermore, reference 27 is a retrospective cohort in which hypoxemia is related with a prolonged length of stay. However, hypoxemia could also be an intermediate for underlying diseases that were not measured or known.

REVIEWER	Emilie Belley-Côté McMaster University, Canada
REVIEW RETURNED	18-Aug-2016

GENERAL COMMENTS	Very interesting research question that is clinically relevant and patient-important. The results are interesting and definitely warrant publication. Here are my comments/suggestions: Please spell PSA in the abstract More details about methods would be useful in the abstract. For example: duplicate screening, eligibility assessment, data extraction, ROB assessment?

	It may be useful to define respiratory compromise in the abstract as it was not clear to me what it included. Consider adding a sentence in the background explaining how your systematic review adds to the one previously done (discussed in the abstract). According to the last sentence of the background, maintaining consistent definitions of adverse events was emphasized. However, I cannot find these definitions in the methods. Also, from previous work on SRs, I know how definitions tend to vary from study to study. How was that handled? The search was conducted over a year ago. Would it be worth updating to see if there are new eligible studies. Why did you choose to use the Jadad score to assess ROB. Cochrane advises against the use of tools providing scores. Moreover, the Jadad score with the additional criteria may need to be validated. In the future, you may want to use validated filters to selects RCTs (the SIGN filters for example) rather than adding RCT as a Mesh term and free text to your search strategy. Why did you exclude trials that included less than 40 patients in either arm? The "sensitivity analyses" that were performed are very similar to subgroup analyses. Should they be referred to as such? It may be interesting to report on your a priori hypotheses. Were these analyses planned a priori? The results are reported as Odds Ratios. However, odds ratios are harder to interpret for the clinicians. I suggest reporting Risk Ratios if possible. Also, in page 8 line 25, you interpret your OR like a RR which is incorrect. If possible, use the Cochrane classification of heterogeneity: might not be important, moderate, substantial, considerable. Please reference the studies in statements such as "six studies reported". Random-effects and fixed-effects analyses are expected to yield similar results, but the random-effects model usually gives wider confidence intervals. I suggest a conservative approach reporting only random-effects analyses, that would make the results section lighter. Page 10-line 7, the need to provide Page 10 - line 10, I wonder if the low number of observed events and the subsequent lack of power would not be a simpler statement. Page 10 line 39, outcomes differed Although the discussion about optimal information size in the last paragraph of page 10 is quite interesting, I wonder if it does not

	belong in the discussion. Page 12 - line 42, I suggest updating the meta-analysis to include that trial.

REVIEWER	Gowri Raman Tufts Medical Center, Boston, MA, USA
REVIEW RETURNED	19-Aug-2016

GENERAL COMMENTS	Overall, this is a well-written manuscript examining patient safety during procedural sedation using capnography. The study is conducted as a systematic review and meta-analyses of randomized controlled trials. Specific suggestions related to methods and statistical analyses follow: 1) In methods section, authors should acknowledge that they did not search for unpublished literature and they did not conduct any statistical analyses to assess publication bias. 2) Under the quality and potential bias section, authors have added declaration of funding source and conflicts of interest to the modified Jadad quality score. Please clarify if this is a validated scoring method of risk of bias and also, suggest providing a reference if this method has been used elsewhere. 3) Throughout the results section, odds ratio has been interpreted as "incidence" and should be appropriately revised. 4) On page 8, authors report requiring at least 3 studies to conduct meta-analysis. Contrary to this statement, a meta-analysis was performed with 2 studies (please see figure 7A) 5) On page 10 authors acknowledge lack of significance owing to the low number of observed events, but chose to use M-H random effects meta-analysis. In view of low number of observed events, authors should use Peto method for this meta-analysis. 6) It is unclear what the power calculation of mortality/mobidity adds to this paper, since mortality due to cardiac dysrhythmias or arrest are not reported in the results section. 7) Figure 1 A and 5 A are the same in terms of the studies and results; and it is unclear why this figure appears twice. 8) Authors have combined studies conducted in emergent/urgent care setting with studies conducted in elective setting. A sensitivity analysis can be used for this comparison. 9) Suggest rechecking of Langham et al 2015 that has no effect on outcomes (odds ratio 1.0) in Figures 1 and 5. This study result is in contrast to all other studies included in the analyses. 10) In table 1, it is unclear where the scoring for potential for bias came from. Please clarify the difference between low: 0 and No: 0.

VERSION 1 – AUTHOR RESPONSE

Reviewer: 1

Reviewer Name
Melissa Langhan

Institution and Country
Associate Professor
Yale University

New Haven, CT, USA

Please state any competing interests or state 'None declared':
None declared

Please leave your comments for the authors below

This was a very well written manuscript. My only comment would be to consider clarifying the primary endpoints - they are not well defined in the methods, and with a small number of eligible studies, it may be feasible to contact the authors for more information on specific endpoints to improve homogeneity of results.
Thank you to the reviewer for their kind comments. In response to their request, we have included the exact mild/severe desaturation endpoints of all of the included studies are in Table 5. We have also updated the endpoints in the methods to define severe desaturation as being *"defined as SpO2 ≤85%".*

Reviewer: 2

Reviewer Name
Kim van Loon

Institution and Country
University Medical Center Utrecht, The Netherlands
PhD student and specialist registrar in Anaesthesia

Please state any competing interests or state 'None declared':
None declared

Please leave your comments for the authors below
The authors of this manuscript performed a systematic review and meta-analysis to answer the research question whether adding capnography to pulse oximetry and visual assessment would reduce sedation related events during procedural sedation. First I would like to congratulate the authors with the manuscript, it is well written and gives a comprehensive overview of RCT's in this field. However, I also have a few major- and some minor comments that I would like to share with the authors.

1. On page 6, the authors state that they use the modified Jadad score to assess the study quality. The score was developed to evaluate quality of clinical reports in pain relief, thus for intervention studies evaluating treatment. In the Jadad score, double blinding is an important quality indicator. However, intervention studies in this setting compare two monitoring strategies, not therapy alone. Meaning they do not only assess the additional monitor, but also the alarm strategy, how the caregivers respond and what they do (therapy) with the additionally received information. In real life, capnography readings are interpreted by the caregivers and not by an independent observer in a semi-experimental setting. The interpretation of multiple monitors, that sometimes give contradictory information, can be demanding and confusing. Therefore, I consider the quality indicators chosen not fully applicable to intervention studies evaluating monitoring strategies. This is consistent with the reasoning of Conway et al. (p12 line 25-45) I would suggest modifying the quality score or alternatively describing this more clearly as a limitation in the discussion.
Thank you to the reviewer for their comments. In response, we have added a paragraph to the discussion clarifying the use of the Jadad score and the justification of the inclusion of double blind studies.

As Kim has noted, this perspective raised by Conway et al. is also an area that we considered as part of the design of our analysis. Indeed, we have assumed the viewpoint that there are two types of evidence relevant to medical decision making, (1) randomized controlled trials and (2) real-world evidence. It is well known that populations enrolled in trials do not always match well to those of populations exposed to an intervention in real life. Furthermore, in 2012 Veerus et al. showed (in "Results from a blind and a non-blind randomised trial run in parallel: experience from the Estonian Postmenopausal Hormone Therapy (EPHT) Trial") that results from parallel trials (one blinded and one not) could differ. As such, blinded trials (although maybe not representative of real world practice) are an important method of comparing two interventions without high risk of operator bias. It is for this reason that we chose to use the Jadad score. As monitoring is, as the reviewer noted, a special case in trial design we also adapted this score to make it more applicable. Under standard protocols for systematic reviews, a scoring system for study quality is required. The most commonly used score is the Jadad score. Other scores include Delphi, Consort, and Cochrane, all of which considered blinding/masking to be an important part of trial quality (see Berger 2009: A General Framework for the Evaluation of Clinical Trial Quality). As such, highlighting the use of a scoring system as a limitation in our analysis would seem counter to accepted practice.

2. In the sensitivity analysis the authors add 7 scenarios after they have evaluated several outcomes (e.g. mild and severe hypoxemia, apnoea) with subgroups of endpoint definitions and study qualities. That adds up to a lot of comparisons that could potentially result in a type I error. As the authors performed mixed effect models, they could consider, when they consider these scenarios (p7 line 23-30) as important confounders, to add them as fixed effects in the mixed models.
We appreciate the reviewer's suggestion. We have added the following text to the manuscript methods section to clarify this: *"No formal statistical comparisons were made between sensitivity analyses, and intervention effects were not calculated for the excluded studies, thereby mitigating the introduction of type 1 error into the analysis."*

The analyses conducted are distinct from subgroup analyses. We have consulted the Cochrane Handbook on sensitivity analysis and have confirmed the following (http://handbook.cochrane.org/chapter_9/9_7_sensitivity_analyses.htm);

> *"Sensitivity analyses are sometimes confused with subgroup analysis. Although some sensitivity analyses involve restricting the analysis to a subset of the totality of studies, the two methods differ in two ways. First, sensitivity analyses do not attempt to estimate the effect of the intervention in the group of studies removed from the analysis, whereas in subgroup analyses, estimates are produced for each subgroup. Second, in sensitivity analyses, informal comparisons are made between different ways of estimating the same thing, whereas in subgroup analyses, formal statistical comparisons are made across the subgroups."*

Finally, we would also note that our primary analysis was conducted using a purely random effects model rather than a mixed effects model (which would capture both fixed and random effects). In turn, we believe the inclusion of fixed effects to mitigate type 1 error would not be feasible without substantially adjusting the methodology employed in the analysis.

3. I would consider "the administration of standard vs. rescue supplemental oxygen" as an important confounder for the determinant-outcome relation (the research question) as stated on p4 line 48-55. I would suggest adding this characteristic in table 1, and if possible also as a fixed effect in the mixed models (I could imagine that the model is not able to converge with too many fixed effects). Furthermore, the characteristic "who was responsible for monitoring the patient (e.g dedicated nurse, doctor)" would be a nice complement to table 1.
We agree with the reviewer, and have added both suggested items to Table 1.

Minor comments:

- The abstract, p2 line 10, PSA is not fully written in the abstract.
We have rectified this in the revised version of the manuscript.

- p4, line 11, PSA definition could also be considered for surgical procedures, not provided only for non-surgical procedures.
We have clarified that PSA may also be used for minor surgical procedures.

- p5, line 24, you could consider using Web of Knowledge to assure that you are not missing any important papers by screening the references and citations of the included papers.
We have clarified in the methods section that reference lists in the included papers were not searched in addition to the PubMed, Cochrane Library, and EMBASE database searches.

- p12, line 57, I would refrain from using reference 26. The paper of Ehrenfeld et al. discussed hypoxemia in the surgical population, a completely different setting and population from the PSA population. Furthermore, reference 27 is a retrospective cohort in which hypoxemia is related with a prolonged length of stay. However, hypoxemia could also be an intermediate for underlying diseases that were not measured or known.
We have addressed this by making a number of clarifying changes to paragraph 8 of the discussion.

Reviewer: 3

Reviewer Name
Emilie Belley-Côté

Institution and Country
McMaster University, Canada

Please state any competing interests or state 'None declared':
None declared

Please leave your comments for the authors below
Very interesting research question that is clinically relevant and patient-important. The results are interesting and definitely warrant publication.
We thank the reviewer for their commendation.

Here are my comments/suggestions:
Please spell PSA in the abstract
We have rectified this in the revised version of the manuscript.

More details about methods would be useful in the abstract. For example: duplicate screening, eligibility assessment, data extraction, ROB assessment?
We have added the following clause to the abstract: *"screening and data extraction was conducted by two independent reviewers, and study quality was assessed using a modified Jadad scale".*

It may be useful to define respiratory compromise in the abstract as it was not clear to me what it included.
We have added the following parenthetic clarification to the abstract: "poor or absent ventilation cascading into hypoxia, tissue injury and cardiac decompensation"

Consider adding a sentence in the background explaining how your systematic review adds to the one previously done (discussed in the abstract).
We have added the following to the methods-endpoints section: "In contrast to the previous meta-analysis,16 our analysis includes sufficient data to examine individual endpoints, e.g. minor

desaturation, as opposed to composite endpoints (e.g. desaturation, apnea, or respiratory depression). We also examine data by defined endpoints, such as oxygen desaturation <90% and <85%, which was not previously possible."

According to the last sentence of the background, maintaining consistent definitions of adverse events was emphasized. However, I cannot find these definitions in the methods. Also, from previous work on SRs, I know how definitions tend to vary from study to study. How was that handled?
We thank the reviewer for bringing this to our attention and we hope that updates described previously have made our approach more transparent. As previous analyses have created composite endpoints, we focused on consistent endpoints (e.g. mild desaturation) and where possible prevented analyses of trials that reported on exactly the same endpoint, e.g. oxygen desaturation <90%.

The search was conducted over a year ago. Would it be worth updating to see if there are new eligible studies.
We appreciate the reviewer's concerns but remain hopeful the 2015 cut-off date isn't considered a hindrance to publication of the work.

Why did you choose to use the Jadad score to assess ROB. Cochrane advises against the use of tools providing scores. Moreover, the Jadad score with the additional criteria may need to be validated.
Thank you to the reviewer for posing this question. We have added a sentence to the end of the "Quality and potential bias" section to clarify this.

The Jadad score was used to assess study quality rather than risk of bias. Bias was instead assessed using a manufacturer funding/conflict of interest assessment, which is based on systematic reviews such as Lexchin *et al.* (BMJ. 2003;326(7400):1167–70), which demonstrate systematic bias in favor of products which are made by the company funding the research. Crucially, we do not make the deductive fallacy that the absence of manufacturer funding indicates an absence of bias, but the presence of funding from industry is a strong indicator of bias. Indeed, the Cochrane Collaboration has acknowledged that such bias cannot be detected using the Cochrane risk of bias tool.

In the future, you may want to use validated filters to selects RCTs (the SIGN filters for example) rather than adding RCT as a Mesh term and free text to your search strategy.
We thank the reviewer for bringing this to our attention. We are aware of a number of potential approaches to identifying RCTs and selecting an appropriate set of search terms is challenging. However, we note that "Randomized Controlled Trial"[Publication Type] alone has a reported sensitivity of 93.7% (McKibbon *et al.* Health Information & Libraries Journal. 2008;26:187–202.). The addition of free-text terms and "Randomized Controlled Trials as Topic" would likely improve the sensitivity of the search to higher than that of the Marson 2 strategy (Nwosu *et al.* Obstetrics and Gynecology. 1998;91:618–22) at 96.8% which we could consider to be sufficiently high for the present study.

Why did you exclude trials that included less than 40 patients in either arm?
The exclusion of trials enrolling fewer than 40 patients per arm was part of the analysis protocol. It was introduced because the small sample sizes can make interpretation of data difficult. The value of 40 was obtained by using the sample size calculation equation for statistical superiority design from Zhong 2009, "How to Calculate Sample Size in Randomized Controlled Trial?" and data from Qadeer et al. 2009 (Capnographic monitoring of respiratory activity improves safety of sedation for endoscopic cholangiopancreatography and ultrasonography). If 31% of patients complete without hypoxemia using standard of care and 54% complete without hypoxemia using capnography (as reported by Qadeer et al), then the trial size must be over 71 patients. For the purposes of the present study, this 80, or 40 per arm.

The "sensitivity analyses" that were performed are very similar to subgroup analyses. Should they be referred to as such? It may be interesting to report on your a priori hypotheses. Were these analyses planned a priori?
Regarding the point on subgroup analyses, we have added two sentences to the manuscript: *"No formal statistical comparisons were made between sensitivity analyses, and intervention effects were not calculated for the excluded studies, thereby mitigating the introduction of type 1 error into the analysis."* and a sentence noting that the analyses were specified *a priori*.

The results are reported as Odds Ratios. However, odds ratios are harder to interpret for the clinicians. I suggest reporting Risk Ratios if possible. Also, in page 8 line 25, you interpret your OR like a RR which is incorrect.
We have rectified the odds ratio error in the text, furthermore we have added the risk ratio results for base case analyses to Table 2.

If possible, use the Cochrane classification of heterogeneity: might not be important, moderate, substantial, considerable.
Thank you for this suggestion. In response, we have updated the manuscript to include classifications of heterogeneity based on Higgins *et al.* BMJ. 2003;327(7414):557–560.

Please reference the studies in statements such as "six studies reported".
The corresponding references have been added to the sections of the manuscript which report the number of studies reporting each endpoint.

Random-effects and fixed-effects analyses are expected to yield similar results, but the random-effects model usually gives wider confidence intervals. I suggest a conservative approach reporting only random-effects analyses, that would make the results section lighter.
We have removed the reporting of fixed effects outcomes from the results section.

Page 10-line 7, the need to provide
This has been corrected in the updated version of the manuscript.

Page 10 - line 10, I wonder if the low number of observed events and the subsequent lack of power would not be a simpler statement.
This has been changed in the updated version of the manuscript.

Page 10 line 39, outcomes differed
This has been corrected in the updated version of the manuscript.

Although the discussion about optimal information size in the last paragraph of page 10 is quite interesting, I wonder if it does not belong in the discussion.
The discussion has been moved in the updated version of the manuscript.

Page 12 - line 42, I suggest updating the meta-analysis to include that trial.
We appreciate your question, but in order to maintain the systematic nature of the review it is, unfortunately, not possible to include single studies identified *post-hoc*.

Reviewer: 4

Reviewer Name

Gowri Raman

Institution and Country

Tufts Medical Center, Boston, MA, USA

Please state any competing interests or state 'None declared':

None

Please leave your comments for the authors below

Overall, this is a well-written manuscript examining patient safety during procedural sedation using capnography. The study is conducted as a systematic review and meta-analyses of randomized controlled trials.

Specific suggestions related to methods and statistical analyses follow:

1) In methods section, authors should acknowledge that they did not search for unpublished literature and they did not conduct any statistical analyses to assess publication bias.
We have updated the methods to include the following item in response to the reviewer's comment: "*No "grey" or unpublished literature was included in the search strategy and, as the review protocol was not registered in advance, the full search strategy (Supplement, Table 3) and additional details are provided in the Supplement."*

2) Under the quality and potential bias section, authors have added declaration of funding source and conflicts of interest to the modified Jadad quality score. Please clarify if this is a validated scoring method of risk of bias and also, suggest providing a reference if this method has been used elsewhere.
We have clarified in the second paragraph that the risk of bias assessment was performed separately from the quality assessment (as opposed to having been integrated into the modified Jadad score).

3) Throughout the results section, odds ratio has been interpreted as "incidence" and should be appropriately revised.
We thank the author for identifying this and have made updates in the text accordingly.

4) On page 8, authors report requiring at least 3 studies to conduct meta-analysis. Contrary to this statement, a meta-analysis was performed with 2 studies (please see figure 7A)
We have added a note on page 8 to clarify this. As apnea was included in our pre-defined list of endpoints, it was included in the meta-analysis although only two high-quality studies were identified. The cut-off of three studies was, as specified in the methods, for other endpoints not pre-defined in our protocol but identified during the review.

5) On page 10 authors acknowledge lack of significance owing to the low number of observed events, but chose to use M-H random effects meta-analysis. In view of low number of observed events, authors should use Peto method for this meta-analysis.
We thank the reviewer for the recommendation and have calculated the Peto odds ratio and included it in the updated manuscript.

6) It is unclear what the power calculation of mortality/mobidity adds to this paper, since mortality due to cardiac dysrhythmias or arrest are not reported in the results section.
Most studies in PSA to date have focused on more frequent but less severe events, such as oxygen desaturation. This has proved controversial due to the differing opinions on how relevant these events are to patient safety outcomes. As we have experienced in debates taking place at international congresses, physicians undertaking PSA would like to know whether use of capnography will impact on mortality and major morbidity. Our calculations indicate that for an RCT, this is likely not feasible to show with significance.

7) Figure 1 A and 5 A are the same in terms of the studies and results; and it is unclear why this figure appears twice.
This has been rectified in the updated manuscript.

8) Authors have combined studies conducted in emergent/urgent care setting with studies conducted in elective setting. A sensitivity analysis can be used for this comparison.
We appreciate the reviewer's comment and during our study found that analysis in certain sub-populations was appropriate. We have presented our findings in other forms. For example, a talk on those studies related to gastrointestinal procedures was presented at Digestive Disease Week 2016 and is available online.

9) Suggest rechecking of Langham et al 2015 that has no effect on outcomes (odds ratio 1.0) in Figures 1 and 5. This study result is in contrast to all other studies included in the analyses.
We can confirm that Table 1 in Langhan *et al.* reported 23 desaturation events in each arm. No changes have therefore been made to the manuscript in response to this comment.

10) In table 1, it is unclear where the scoring for potential for bias came from. Please clarify the difference between low: 0 and No: 0.
Thank you for this suggestion, we have now changed the "No's" (which originated in a previous draft) to Low.

We very much hope that the above changes satisfactorily address the concerns raised by the reviewers and that the resulting manuscript is of a sufficiently high standard for the esteemed readership of *BMJ Open*.

VERSION 2 – REVIEW

REVIEWER	Emilie Belley-Cote McMaster University, Canada
REVIEW RETURNED	08-Nov-2016

GENERAL COMMENTS	Even though the question addressed remains relevant and the paper is well written, the authors have failed to address some major concern. I am especially concerned by the relevance of this systematic review in the current literature when the authors acknowledge the publication of studies meeting inclusion criteria since they conducted their search. page 2 line 30 - were instead of was page 3 1st bullet - I think it might be an overstatement to say that it provides consistent evidence of improvements in patient safety I still think it would be worth reviewing the grey literature as the industry might have performed some studies that were never published in indexed journals. Potentially leading to publication bias which could not be truly assessed visually or statistically given the number of studies reporting most outcomes. I am concerned that rather than making the Jadad score more valid by modifying it, the authors are transforming a validated score, albeit with serious limitations, into a score of unknown significance. Reporting the results as RR rather than OR would make them easier to understand for knowledge users and only requires clicking on a box in RevMan. I wish the authors had updated their search to 2016. Especially since they are aware of studies published since they conducted the

	search. Also, it is indeed possible to include studies identified through means other than the database search strategy. One of the purposes of systematic reviews is to provide sufficient power to assess for rare outcomes for which individual studies may be underpowered. By excluding studies deemed too small, the authors limit the power of their analyses without any benefit to the validity of the meta-analysis' results.

REVIEWER	Gowri Raman Tufts Medical Center, Boston, MA, USA
REVIEW RETURNED	27-Nov-2016

GENERAL COMMENTS	Thank you for revising per reviewer's suggestion. The revised manuscript reads well. Here are few minor suggestions to enhance transparency of the methodology: 1) In the abstract section, lines 26-28 mention search was "2015(inclusive)." The next sentence indicates that it was only conducted until June 2015. Please edit the 2015 (inclusive) to indicate it was only until June 2015. 2) Page 5, lines 39-41 It will be helpful for readers to know the reason regarding search start date of 1995. 3) Page 10, lines 14-31, heterogeneity measure I-square is missing for summary odds ratio in the section on assisted ventilation 4) Table 1, column Modified Jadad score will require a footnote for either "higher scores indicate high quality studies" or 5.5–8 were designated as high quality studies. The next column potential for bias will require a footnote that this score is derived independent of modified Jadad score through the declaration of funding sources and conflicts of interest. These are needed for couple of reasons: First, the two scores are in reverse direction (high scores are high quality in modified Jadad score and low scores mean low biases in the next column); Second, both measures do not appear to correlate well (studies of high quality were deemed to have moderate or high biases e.g, Beitz et al and Qadeer et al and studies of poor quality were deemed to have low biases e.g, Kochhar 2015 and Mehta 2014) .

VERSION 2 – AUTHOR RESPONSE

Reviewer: 3

Reviewer Name

Emilie Belley-Cote

Institution and Country

McMaster University, Canada

Please state any competing interests or state 'None declared':

None declared

Please leave your comments for the authors below

Even though the question addressed remains relevant and the paper is well written, the authors have failed to address some major concern. I am especially concerned by the relevance of this systematic review in the current literature when the authors acknowledge the publication of studies meeting inclusion criteria since they conducted their search.

We fully appreciate the consistent encouragement of the reviewer to ensure our meta-analysis is strengthened by reflecting all currently available studies. Our newly revised manuscript now reflects the results of rigorous methodology that was required to address this comment by re-running the literature searches through December 31, 2016. The searches were dual-screened by RS and RFP and ultimately resulted in the addition of Riphaus *et al.* 2016 and Klare *et al.* 2016, and the consolidation of the Kochhar *et al.* 2015 and Mehta *et al.* 2014 studies (both previously in abstract form) into a single full-publication (Mehta *et al.* 2016). Dual data extraction was then conducted on the three new studies, all analyses have been re-run to incorporate new data and the manuscript and supplementary information have been updated in line with the new findings.

page 2 line 30 - were instead of was

We thank the reviewer for bringing this to our attention.

page 3 1st bullet - I think it might be an overstatement to say that it provides consistent evidence of improvements in patient safety

We have modified the wording to note that the modern nature of the studies provides “clinical relevant” evidence.

I still think it would be worth reviewing the grey literature as the industry might have performed some studies that were never published in indexed journals. Potentially leading to publication bias which could not be truly assessed visually or statistically given the number of studies reporting most outcomes.

The EMBASE searches conducted as part of the original review and the update contained numerous abstracts, including the Kochhar and Mehta studies that were included the original manuscript (and have been subsequently replaced by the consolidated Mehta *et al.* 2016 manuscript). There are, of course, well known logistical issues associated with extracting data and performing a suitably rigorous bias and quality appraisal from the 250-300 words that are typically permitted in abstract submissions. Nevertheless, we have clarified in our manuscript that the original searches did include abstracts (and that they were not excluded per protocol).

I am concerned that rather than making the Jadad score more valid by modifying it, the authors are transforming a validated score, albeit with serious limitations, into a score of unknown significance.

We have expanded the discussion on the limitation of using a modified Jadad score, but would note that our mixed model analysis with either the Jadad score or modified Jadad score as covariates illustrated that there is a strong ($R2 = 0.93$) correlation between the two measures and that the score is therefore not of unknown significance. We do continue to state that the nature of the blinding options available in capnography trials make an assessment using existing tools problematic in different ways. We hope that the expanded acknowledgement of the limitations associated with the modifications (and the analysis thereof) are sufficient to allay the reviewer’s concerns.

Reporting the results as RR rather than OR would make them easier to understand for knowledge users and only requires clicking on a box in RevMan.

We have updated the manuscript to incorporate both odds ratios and relative risks in the manuscript text and Table 2 has been updated to report exclusively relative risks. The reason for this is that the feedback from reviewers as to which is most pertinent varies.

I wish the authors had updated their search to 2016. Especially since they are aware of studies published since they conducted the search. Also, it is indeed possible to include studies identified through means other than the database search strategy.

We appreciate the reviewer's wish and hope that the search updates and extensive re-analysis address this concern.

One of the purposes of systematic reviews is to provide sufficient power to assess for rare outcomes for which individual studies may be underpowered. By excluding studies deemed too small, the authors limit the power of their analyses without any benefit to the validity of the meta-analysis' results.

We appreciate this concern and the thinking underpinning it. In response, we note the rationale that is included in the supplementary information on the powering calculation. Extensive research also shows the differences in outcomes in "small" versus "large" trials (Nüesch *et al.* BMJ 2010;341:c3515, Zhang *et al.* Crit Care. 2013; 17(1): R2, Farkouh and Fuster. Nat Clin Pract Nephrol. 2008;4(3):115, etc.) with larger trials typically yielding more conservative outcomes (i.e. smaller trials tend to overestimate the effect size). It is with these cautionary studies in mind that we omitted smaller studies from the present analysis.

Reviewer: 4

Reviewer Name

Gowri Raman

Institution and Country

Tufts Medical Center, Boston, MA, USA

Please state any competing interests or state 'None declared':

None declared

Please leave your comments for the authors below

Thank you for revising per reviewer's suggestion. The revised manuscript reads well. Here are few minor suggestions to enhance transparency of the methodology:

1) In the abstract section, lines 26-28 mention search was "2015(inclusive)." The next sentence indicates that it was only conducted until June 2015. Please edit the 2015 (inclusive) to indicate it was only until June 2015.

We thank the reviewer for their keen eye on this. Other updates made in response to comments from other reviewers have mitigated this concern as the updated searches now cover up to the end of 2016.

2) Page 5, lines 39-41 It will be helpful for readers to know the reason regarding search start date of 1995.

We have clarified that a previous meta-analysis did not identify any studies on capnography versus standard of care published prior to 1995.

3) Page 10, lines 14-31, heterogeneity measure I-square is missing for summary odds ratio in the section on assisted ventilation

Heterogeneity measures have been added to this section of the manuscript.

4) Table 1, column Modified Jadad score will require a footnote for either "higher scores indicate high quality studies" or 5.5–8 were designated as high quality studies. The next column potential for bias will require a footnote that this score is derived independent of modified Jadad score through the declaration of funding sources and conflicts of interest. These are needed for couple of reasons: First, the two scores are in reverse direction (high scores are high quality in modified Jadad score and low scores mean low biases in the next column); Second, both measures do not appear to correlate well (studies of high quality were deemed to have moderate or high biases e.g, Beitz et al and Qadeer et al and studies of poor quality were deemed to have low biases e.g, Kochhar 2015 and Mehta 2014).

We have clarified both of these aspects of the bias and quality assessments in the footnotes immediately beneath the table.

We very much hope that the above changes satisfactorily address the concerns raised by the reviewers and that the resulting manuscript is of a sufficiently high standard for the esteemed readership of *BMJ Open*.

VERSION 3 – REVIEW

REVIEWER	Emilie Belley-Cote McMaster University
REVIEW RETURNED	06-Mar-2017

GENERAL COMMENTS	Great improvement in the manuscript. Thank you for updating the search. It is a bit unusual to report both RR and OR, I suggest reporting only RR. Also for assisted ventilation on P10-11, only ORs are reported. The authors state that calculating optimal information size was their second objective, but they report on that only in the discussion section (as opposed to reporting with the results). I suggest removing as an objective and only discussing in the discussion.

VERSION 3 – AUTHOR RESPONSE

We are very pleased with your recommendation for publication and would like to thank the reviewer for her additional comments to improve the manuscript. As requested, we have now removed the power calculation from the study objectives. The power calculation is now introduced and discussed in the discussion section only. The risk ratio (RR) and odds ratio (OR) comment is more complex. First, we respectfully note that we had previously received comments in this review process favouring RR as the main outcome. In addition, due to constraints involved with analysis of endpoints with low incidence (<1%), we have utilized the Peto method, which by design provides only an OR. To more easily allow for comparison between all endpoints, we have therefore included both the RR and OR throughout. We have clarified this technical issue in the methods section.