PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	The Quality of Reporting of Pilot and Feasibility Cluster Randomised Trials: A Systematic Review
AUTHORS	Chan, Claire; Leyrat, Clémence; Eldridge, Sandra

VERSION 1 - REVIEW

REVIEWER	Ruth Pickering University of Southampton UK
REVIEW RETURNED	28-Apr-2017

GENERAL COMMENTS	This is an interesting and well written paper reporting an extension of previous work reviewing the content of published feasibility/pilot trials in relation to individually randomised trials, to feasibility/pilot trials relating to cluster randomised trials. I have the following points to raise. 1 Page 5, paragraph starting on line 42, describing the items on the review form and where they came from. The source of items seems to be important, and in the Discussion the authors mention that items from some sources were better addressed than other items. I found the explanation difficult to follow. If the source is important, perhaps it could be indicated in a new column in Table 4. 2 Page 10, line 20/21, the advice on reporting pilot/feasibility trials is clear that effectiveness/efficacy results shouldn't be included, yet they are widely reported (76% of the papers in the current review). Some thoughts: is there any value in checking out the statistical analysis planned for the main trial, some assumptions may not be met, transforming outcome or covariates for example, is it that the pilot study is likely to be too small to check any assumptions, or that if such checking is done it shouldn't be published? If the pilot study fails to show any indication that the intervention is beneficial (or perhaps a trend towards it being detrimental) should this not be allowed to impact on the decision as to whether to continue to a definitive trial? Does it matter if this was or wasn't prestated. 3 Page 10, line 55, sample size rationale. They refer to sample size rationale at a number of place, and I felt it sometimes wasn't clear whether they were talking about the sample size for the pilot/feasibility trial or that for the future definitive trial. On Page 12, line 2/3, they say just 17% of the studies considered the cluster design in the sample size rationale. I did wonder how the authors of this review expect it to be considered. On line 10, the say including a number of clusters with different characteristics would inform on implementation across different clusters: but this statement is more about the type of cluster to include not the number of clusters. On

	page 11, lines 43-45 they say Arain et al reported 36% of studies performed sample size calculations (it isn't clear from this whether it relates to the sample size for the pilot or the future definitive trial), and 17% performed calculations in the current review relating to a feasibility objective. So from this it sounds like a calculation of sample size for the pilot trial may be expected, not just a discussion of selection criteria for the clusters or the statement of a target number. Perhaps it would be helpful to quote sentences from one of the reviewed trials where a good example of a sample size justification/calculation was reported, to give us a clearer idea of what is required. I felt there were a number of the items that would be understood better if the authors could give examples of good description/coverage maybe in comparison to unsatisfactory description. 4 Page 11, line 3, they report tables showing that a table of cluster characteristics was one of the items that was not well reported. In table 2, they report that one of the studies involved only 2 clusters, and that the IQR went from 4 to 16, so that approx. 25% had 4 or fewer. Would it be sensible to include a table describing just 2 clusters or even 4? 5 Page 11, lines 24, the item on unintended consequences not being well reported. In the data extracted Appendix table they explain this relates to unexpected findings that were not a prestated objective and that might have consequences for the design of a definitive trial. On Page 10, lines 39-49, they discuss that the reasons for a pilot trial being conducted, and the progression criteria, being poorly reported, and that sticking to the checklist guards against selective reporting. Doesn't the item on unintended consequences specifically allow selective reporting of unanticipated problems that might impact on the design of a future trial? As these are unexpected consequences and so not prestated, authors have to select them for inclusion in their paper. 6 Page 4, line 9, the inclusion criterion, that the pilot trial was in preparation for a trial assessing effectiveness/efficacy. Can they be more specific here as to whether they mean a subsequent trial conducted by the authors of the pilot trial, or a subsequent trial by anyone, ie the pilot trial potentially presenting useful information for unspecified future researchers to design a trial. On page 12, 35-37, it seems that most of the reviewed trials were aimed at providing information for other researchers or raising questions for future research, rather than being in preparation for a trial assessing effectiveness/efficacy. I got the impression from this, that the authors of the review feel that a pilot trial should be in preparation for a future trial conducted by the authors of the pilot trial, but I don't think they explicitly state this. Does it invalidate a pilot trial, if the information it provides is aimed at generally informing researchers in the area who may be considering doing a trial? 7 In Table 4 the % of reviewed pilot studies adhering to each of the items on the review form, the pilot quality criteria, are reported. Perhaps this last point was covered in the description of the selection of items, on page 5, line 42, which I found difficult to follow. Some feasibility studies have specific and limited objectives, is it possible to blind participants for example, or can recruited participants be retained. If this were the case, would all of the criteria be necessary?

REVIEWER	Dr Jennifer Lewis ScHARR University of Sheffield United Kingdom
REVIEW RETURNED	11-May-2017

GENERAL COMMENTS	The authors present a clear and comprehensive review of 18 pilot cluster-randomised controlled trials, and argue compellingly for an improvement in the reporting quality of such trials. The search and review strategy employed is commendably robust, particularly the validation of the screening procedure. Similarly, the compilation and use of a clear checklist for assessing quality based on CONSORT gives a valid and objective reference for the assessment of quality. That only 18 studies were found that met the eligibility criteria is an important finding in itself. The further investigation, point by point, of reporting quality, is concise and informative, and the finding that those few trials that are published are poorly reported, is timely and significant. Generally, this manuscript is of high quality, being well researched and reported. However, there are some points on which I would like to see additional detail. 1. While the manuscript clearly details how many of the papers included each item on the checklist, it would have been useful if there had been some indication of how many checklist items each individual paper included. We cannot tell, at present, if there were any 'gold standard' papers that reported all, or nearly all, of the items, or if all papers were lacking several. I would therefore recommend that each of the papers included in the review is given an overall 'excellence' score showing how many items that paper reported (e.g., one paper reported 90% of items, 3 papers reported 80% of items… etc). Such a score should not be regarded as definitive, since there are of course elements of reporting quality that cannot feasibly be included in CONSORT (or the abridged checklist used here), but it would be extremely useful in the context of finding examples of good reporting. This would also allow the reader to see whether general reporting quality has increased over time as people become more aware of CONSORT extensions. 2. Secondly, the authors should indicate if certain items tend to get reported together, or not at all, which might indicate attention to of neglect of whole categories of CONSORT. Alternatively, more haphazard neglect of individual items may reflect a tendency to omit information that is either difficult to obtain for a given type of design, or is being deliberately omitted in order to improve the impression of the study. The authors do touch on this in their discussion, which indicates that in general, new and substantially adapted items were less well reported, but this could be elaborated: do all papers report one or two of these items, or do a few report all and most report none? Or something else? Identifying patterns here may help the reader understand the implications of the findings and lead to improved reporting. 3. I have some reservations about the small number of papers that were thoroughly reviewed. There are several reasons why such a

	limited number may have been found: • The authors have used a CONSORT criterion ('pilot' or 'feasibility' in title) to find papers which have then been assessed for CONSORT items. While they also include this for the abstract, broadening the search a little, this strategy will almost certainly have resulted in a skewed sample of papers that have a greater tendency to adhere to CONSORT guidelines. • The use of conditions #3, #5 and #6 in concert may have been too restrictive. • The use of only one database, which is a comprehensive but not an exhaustive one, is likely to have missed some eligible papers. • Additional criteria applied during the eligibility assessment has also excluded many papers. While the strategy, restrictions and eligibility criteria are all sensible and there are clear reasons given for them, it is likely that the authors have overestimated the quality of reporting, possibly by quite some margin. Though this is acknowledged, I would like to see some discussion of the implications of this; if the academic community is taking the excluded 'pilot' trials seriously, understanding the quality of reporting therein is also important, arguably more so, since there are more of them and their prevalence may be contributing to continued poor reporting. In particular, the 32 trials excluded for not assessing feasibility would be of interest. Are they all simply underpowered main trials? If so, why might they be being legitimately published as pilot or feasibility studies? If not, what are they addressing, and to what end? A discussion of these points would be helpful to give a more accurate picture of the actual state of trials reported as pilot or feasibility trials. Minor considerations include: 1. What is the 'written guidance' which was used by CC and CL for data extraction? (p5) 2. The items not included because 'we expected the item would be generally well reported' (p5) should be detailed in some way – at the least, report the number of items not included from each category, or if feasible, include these in an appendix so the reader can understand what was not assessed. 3. Similarly, how many/which items were excluded because they were difficult to extract? (p6) 4. Incomplete reference 9 (p15)

VERSION 1 – AUTHOR RESPONSE

Reviewer: 1
Reviewer Name: Ruth Pickering
Institution and Country: University of Southampton, UK
Please state any competing interests or state 'None declared': None Declared

Please leave your comments for the authors below
This is an interesting and well written paper reporting an extension of previous work reviewing the content of published feasibility/pilot trials in relation to individually randomised trials, to feasibility/pilot trials relating to cluster randomised trials. I have the following points to raise.

Thank you.

1 Page 5, paragraph starting on line 42, describing the items on the review form and where they came from. The source of items seems to be important, and in the Discussion the authors mention that items from some sources were better addressed than other items. I found the explanation difficult to follow. If the source is important, perhaps it could be indicated in a new column in Table 4.

We have updated two paragraphs for further clarity, as well as the footnote in table 4. There are only two sources in table 4, the CONSORT extension for pilot trials and the CONSORT extension for CRTs. We differentiate the two using normal font and bold font, as explained in the footnote.

Line 49-54 (pg5) and 6-20 (pg6)
"To assess reporting quality, we created a list of quality assessment items based on the CONSORT extension for pilot trials. We also looked at the CONSORT extension for CRTs, and incorporated any cluster-specific items into our quality assessment items. Where a CRT item became less relevant in the context of a pilot trial, we did not extract it (e.g. whether variation in cluster sizes was formally considered in the sample size calculation). In addition, where there was a substantial difference between the item for the CONSORT extension for CRTs and that for the pilot trial extension and the items were not compatible, we used the latter item (e.g. focusing on objectives rather than outcomes). We recognised the need to balance comprehensiveness and feasibility. Therefore, where items referred to objectives or methods, we extracted this for the primary objective only. We also did not extract on whether papers reported a structured summary of trial design, methods, results, and conclusions."

Line 45-54 (pg8) and 7-8 (pg9)
"The pilot CRTs in our review are published after the CONSORT 2010 for RCTs but before the CONSORT extension for pilot trials. Therefore, to present data on quality of reporting, we looked at our list of quality assessment items based on the CONSORT extension for pilot trials, and grouped reporting items into three categories (Table 4): (1) items in the CONSORT extension for pilot trials that are new compared to CONSORT 2010 for RCTs, (2) items in the CONSORT extension for pilot trials that are substantially adapted from CONSORT 2010 for RCTs and (3) items in the CONSORT extension for pilot trials that are the same as or have only minor differences from CONSORT 2010 for RCTs, plus items in the CONSORT extension for CRTs."

Line 33-41 (Table 4 footnote) (pg22)
"Item numbers in normal font refer to the item in the CONSORT extension for pilot trials that the quality assessment item is based on.
Item numbers in bold italics refer to the item in the CONSORT extension for CRTs that the quality assessment item is based on.
[N] represents new items in the CONSORT extension for pilot trials compared to the CONSORT 2010 for RCTs.
[S] represents items in the CONSORT extension for pilot trials that are substantially adapted from the CONSORT 2010 for RCTs.
*The CONSORT statements do not include an item 13 but there is a participant flow subheading which strongly recommends a diagram. We therefore reference this subheading as 'item 13' here."

2 Page 10, line 20/21, the advice on reporting pilot/feasibility trials is clear that effectiveness/efficacy results shouldn't be included, yet they are widely reported (76% of the papers in the current review). Some thoughts: is there any value in checking out the statistical analysis planned for the main trial, some assumptions may not be met, transforming outcome or covariates for example, is it that the pilot study is likely to be too small to check any assumptions, or that if such checking is done it shouldn't be published? If the pilot study fails to show any indication that the intervention is beneficial (or perhaps a trend towards it being detrimental) should this not be allowed to impact on the decision as to whether to continue to a definitive trial? Does it matter if this was or wasn't prestated.

We have included the following paragraph in our manuscript to clarify this:

Response Line 53 (pg11) and 7-24 (pg12)
“One may however look at potential effectiveness, for example using an interim or surrogate outcome, with a caveat about the lack of power. Moreover, one may include a progression criterion based on potential effect. If so, Eldridge and Kerry recommend any interpretation of potential effect is done by looking at the limits of the confidence interval, and one should also pay attention to features of the pilot which might have biased the result (for example, convenience sampling of clusters). A positive effect finding excluding the null value would still justify the future definitive trial to estimate the effect with greater certainty, but a negative effect finding excluding the null value (i.e. strongly suggesting harm), or even a finding where the clinically important difference is excluded, might suggest not proceeding. It is good practice to pre-state such progression criteria. Finally, one may use estimates from outcome data, for example, as inputs for the sample size calculation for the future definitive trial. In particular, for pilot CRTs we may be interested in estimating the intra-cluster correlation coefficient (ICC), although we note that the ICC estimate from a pilot CRT should not be the only source for the future definitive trial sample size, because of the large amount of imprecision in a pilot trial. ”

3 Page 10, line 55, sample size rationale. They refer to sample size rationale at a number of place, and I felt it sometimes wasn't clear whether they were talking about the sample size for the pilot/feasibility trial or that for the future definitive trial.

Some parameters estimated in the pilot study can inform the sample size determination of the future definitive trial. However, in this review we focus on the justification of the sample size of the pilot trial, which was, on average, poorly reported. We have clarified what we are referring to in various places in the manuscript:

Response Line 49 (pg9), 38 (pg10), 42 (pg11), 33 and 47 (pg12), 48 (pg13), 14 (pg14), 40 (pg15)
“sample size rationale for the pilot trial”

On Page 12, line 2/3, they say just 17% of the studies considered the cluster design in the sample size rationale. I did wonder how the authors of this review expect it to be considered. On line 10, the say including a number of clusters with different characteristics would inform on implementation across different clusters: but this statement is more about the type of cluster to include not the number of clusters.

We have given further explanation on exactly what we counted as considering the cluster design in the sample size rationale in appendix 2, and gave more explanation and an example within the manuscript too:

Response Line 33-39 (Appendix 2) (pg28)
“We required that the authors show some consideration about clustering during the description of their sample size calculation, even if not formally accounting for it currently but describe during their rationale that they e.g. plan to estimate the design effect in the future definitive trial”

Response Line 17-22 (pg14)
“In pilot trials the rationale for considering the clustered design in deciding on numbers in the pilot may be different, for example, considering the number of degrees of freedom needed within each cluster to estimate a variance. In pilot trials, including a number of clusters with different characteristics may also be important to get an idea about the implementation of an intervention across different clusters.”

On page 11, lines 43-45 they say Arain et al reported 36% of studies performed sample size calculations (it isn't clear from this whether it relates to the sample size for the pilot or the future definitive trial),

We have clarified what we are referring to in the manuscript:

Response Line 48 (pg13)
“sample size calculations for the pilot”

and 17% performed calculations in the current review relating to a feasibility objective. So from this it sounds like a calculation of sample size for the pilot trial may be expected, not just a discussion of selection criteria for the clusters or the statement of a target number. Perhaps it would be helpful to quote sentences from one of the reviewed trials where a good example of a sample size justification/calculation was reported, to give us a clearer idea of what is required. I felt there were a number of the items that would be understood better if the authors could give examples of good description/coverage maybe in comparison to unsatisfactory description.

We have tried to make this clearer on the page where we briefly described the sample size rationales given by the eight pilot CRTs reporting one:

Response Line 6-24 (pg10)
“Pilot trials should always report a rationale for their sample size; this can be qualitative or quantitative, but shouldn’t be based on a formal sample size calculation for effectiveness/efficacy. In this review, the rationales were based on logistics, resources, time, a balance of practicalities and need for reasonable precision, a general statement that it was considered sufficient to address the objectives of the pilot trial, formal and non-formal calculation to enable estimation of parameters in the future definitive trial, and a formal calculation based on the primary feasibility outcome. Of these rationales, good examples include “The decision to include eight apartment-sharing communities was based on practical feasibility that seemed appropriate according to funding and the personal resources available”, as well as “The sample size was chosen in order to have two clusters per randomized treatment and the number of participants per cluster was based on the number of degrees of freedom needed within each cluster to have reasonable precision to estimate a variance.”

4 Page 11, line 3, they report tables showing that a table of cluster characteristics was one of the items that was not well reported. In table 2, they report that one of the studies involved only 2 clusters, and that the IQR went from 4 to 16, so that approx. 25% had 4 or fewer. Would it be sensible to include a table describing just 2 clusters or even 4?

We believe it is still important to include cluster-level characteristics, as well as individual-level characteristics, to give an idea of representativeness, especially since imbalance is more likely when there are fewer clusters. The individual and cluster information could be included in the same baseline table. We include the following sentence in our manuscript:

Response Line 49-54 (pg12)
“Although the number of clusters in a pilot trial is usually small it is still important to, for example, describe the cluster-level characteristics using a baseline table as it may give helpful information important for planning the future definitive trial.”

5 Page 11, lines 24, the item on unintended consequences not being well reported. In the data extracted Appendix table they explain this relates to unexpected findings that were not a prestated objective and that might have consequences for the design of a definitive trial. On Page 10, lines 39-

49, they discuss that the reasons for a pilot trial being conducted, and the progression criteria, being poorly reported, and that sticking to the checklist guards against selective reporting. Doesn't the item on unintended consequences specifically allow selective reporting of unanticipated problems that might impact on the design of a future trial? As these are unexpected consequences and so not prestated, authors have to select them for inclusion in their paper.

We agree. This is an unavoidable issue with pilot trials and we have included something on this in the discussion:

Response Line 28-31 (pg13)
"With the item on unintended consequences, we recognise that investigators are free to choose what they interpret and report as an unintended consequence. We recommend careful thought that all unintended consequences that may affect the future definitive trial are reported."

6 Page 4, line 9, the inclusion criterion, that the pilot trial was in preparation for a trial assessing effectiveness/efficacy. Can they be more specific here as to whether they mean a subsequent trial conducted by the authors of the pilot trial, or a subsequent trial by anyone, ie the pilot trial potentially presenting useful information for unspecified future researchers to design a trial. On page 12, 35-37, it seems that most of the reviewed trials were aimed at providing information for other researchers or raising questions for future research, rather than being in preparation for a trial assessing effectiveness/efficacy. I got the impression from this, that the authors of the review feel that a pilot trial should be in preparation for a future trial conducted by the authors of the pilot trial, but I don't think they explicitly state this. Does it invalidate a pilot trial, if the information it provides is aimed at generally informing researchers in the area who may be considering doing a trial?

We don't specify who should carry out the future definitive trial as during the course of a project there may be turnover of staff. However we do expect that the pilot trial be in preparation for a specific future definitive trial planned to go ahead if the pilot trial suggests it is feasible, rather than just a general assessment of feasibility issues to help researchers in general (although pilot trials may do this as an addition). We clarify this in the manuscript. This specific future definitive trial will usually be carried out by the same trial team, although it would not invalidate a pilot trial if it was carried out by a different set of researchers in the area. However, it would be unusual for a team to conduct a pilot trial for others to carry out the future definitive trial, and obtaining funding might be difficult with no long-term plan. A study that provides information aimed at generally informing researchers in the area, without a specific trial in mind yet, would be more appropriately named as a feasibility study rather than a pilot trial, since a feasibility study asks whether something can be done, should we proceed with it, and if so, how, whereas a pilot study asks the same questions but also has the additional feature that a future study, or part of a future study, is conducted on a smaller scale. [Eldridge et al. (2016) Defining Feasibility and Pilot Studies in Preparation for Randomised Controlled Trials: Development of a Conceptual Framework. PLoS ONE. 11(3): e0150205] However, we are focussing on pilot trials only.

Response Line 16-20 (pg4)
"and show evidence that the study was in preparation for a specific trial assessing effectiveness/efficacy that is planned to go ahead if the pilot trial suggests it is feasible (i.e. not just a general assessment of feasibility issues to help researchers in general, although pilot trials may do this as an addition)."

7 In Table 4 the % of reviewed pilot studies adhering to each of the items on the review form, the pilot quality criteria, are reported. Perhaps this last point was covered in the description of the selection of items, on page 5, line 42, which I found difficult to follow. Some feasibility studies have specific and limited objectives, is it possible to blind participants for example, or can recruited participants be

retained. If this were the case, would all of the criteria be necessary?

We agree that not all of the criteria are always necessary. We address this in the footnote of Table 4 where we point out whether the item is not relevant for specific trials. We have added a sentence in the text to draw the reader's attention to this:

Response Line 11-14 (pg9)
"In the tables, denominators for proportions are based on papers for which the item is relevant. Not all items are relevant for all trials, due to their design, so we highlight where this applies in the table footnotes."

Reviewer: 2
Reviewer Name: Dr Jennifer Lewis
Institution and Country: ScHARR, University of Sheffield, United Kingdom Please state any competing interests or state 'None declared': None declared

Please leave your comments for the authors below The authors present a clear and comprehensive review of 18 pilot cluster-randomised controlled trials, and argue compellingly for an improvement in the reporting quality of such trials. The search and review strategy employed is commendably robust, particularly the validation of the screening procedure. Similarly, the compilation and use of a clear checklist for assessing quality based on CONSORT gives a valid and objective reference for the assessment of quality.

That only 18 studies were found that met the eligibility criteria is an important finding in itself. The further investigation, point by point, of reporting quality, is concise and informative, and the finding that those few trials that are published are poorly reported, is timely and significant.

Generally, this manuscript is of high quality, being well researched and reported. However, there are some points on which I would like to see additional detail.

Thank you very much.

1. While the manuscript clearly details how many of the papers included each item on the checklist, it would have been useful if there had been some indication of how many checklist items each individual paper included. We cannot tell, at present, if there were any 'gold standard' papers that reported all, or nearly all, of the items, or if all papers were lacking several. I would therefore recommend that each of the papers included in the review is given an overall 'excellence' score showing how many items that paper reported (e.g., one paper reported 90% of items, 3 papers reported 80% of items… etc). Such a score should not be regarded as definitive, since there are of course elements of reporting quality that cannot feasibly be included in CONSORT (or the abridged checklist used here), but it would be extremely useful in the context of finding examples of good reporting. This would also allow the reader to see whether general reporting quality has increased over time as people become more aware of CONSORT extensions.

Thank you for this suggestion. We have included another table in our manuscript, Table 5, which shows the number (%) of quality assessment items reported by each study. We provide an overall score, as well as a score by categories of CONSORT. We hope this also covers your point 2 below. Furthermore, we include a paragraph discussing Table 5 within the results of our manuscript, and a comment in the discussion:

Response Line 7-24 (pg 11)

“Quality of reporting – by study
Finally, in Table 5 we present the number (percentage) of quality assessment items reported by each study. We provide an overall score, as well as a score by categories of CONSORT. The quality of reporting varies across studies, with 5 of the pilot CRTs reporting over 65% of the quality assessment items and 2 of the pilot CRTs reporting under 30%. There does not appear to be a trend of reporting quality with time. Five of the studies report 90% or more of the quality assessment items in the ‘discussion and other information’ category, and only two studies report less than 50%. Two of the studies report 100% of the items in the ‘title and abstract and introduction’ category, and five studies report less than 50%. The highest percentage of items reported by a study in the ‘methods’ category is 66% and the lowest is 14%. Similarly, the highest percentage of items reported by a study in the ‘results’ category is 78% and the lowest is 18%. Within studies, the category that is best reported tends to be the ‘discussion and other information’ category (had the highest percentage for 10 of the 18 pilot CRTs).”

Response Line 37-38 (pg 13)
“Within studies, the category that is worst reported is the methods, despite being crucial to allow the reader to judge the quality of the trial.”

2. Secondly, the authors should indicate if certain items tend to get reported together, or not at all, which might indicate attention to of neglect of whole categories of CONSORT. Alternatively, more haphazard neglect of individual items may reflect a tendency to omit information that is either difficult to obtain for a given type of design, or is being deliberately omitted in order to improve the impression of the study.

The authors do touch on this in their discussion, which indicates that in general, new and substantially adapted items were less well reported, but this could be elaborated: do all papers report one or two of these items, or do a few report all and most report none? Or something else? Identifying patterns here may help the reader understand the implications of the findings and lead to improved reporting.

Please see response above.

3. I have some reservations about the small number of papers that were thoroughly reviewed. There are several reasons why such a limited number may have been found:

• The authors have used a CONSORT criterion (‘pilot’ or ‘feasibility’ in title) to find papers which have then been assessed for CONSORT items. While they also include this for the abstract, broadening the search a little, this strategy will almost certainly have resulted in a skewed sample of papers that have a greater tendency to adhere to CONSORT guidelines.

Searching for pilot studies by looking for ‘pilot’ or ‘feasibility’ in the title/abstract is usual practice. We reference Lancaster et al. when explaining our search strategy. This search strategy avoids us identifying lots of non pilots. Moreover we anticipate that there will be few pilot or feasibility studies where the authors have not used either of these terms in the title/abstract. Furthermore, if a study did not include either of these terms in the title/abstract, and yet was a pilot, then we would not expect them to report it as they would have done if they had identified it as a pilot study, so it could be seen as unfair assessment. However, we have included this in our limitations:

Response Line 33-39 (pg14)
“Our inclusion criteria stipulated that papers must have the word ‘pilot’ or ‘feasibility’ in the title or abstract, so we may have missed some pilot CRTs and thus may have overestimated the percentage reporting ‘pilot’ or ‘feasibility’ in the title. This strategy may also have resulted in a skewed sample of papers with a greater tendency to adhere to CONSORT guidelines. However, our review suggests

reporting of pilot CRTs need improving, so our conclusion would remain the same."

• The use of conditions #3, #5 and #6 in concert may have been too restrictive.

As above, this is a common search strategy to identify cluster randomised trials, and we referenced Diaz-Ordaz et al. when explaining our search strategy. However we have included this in our limitations:

Response Line 30-32 (pg14)
"…and the use of conditions #3, #5, and #6 (see Appendix 1) may have been restrictive. Our aim was to get a general idea of reporting issues in the area, though, rather than doing a completely comprehensive search."

• The use of only one database, which is a comprehensive but not an exhaustive one, is likely to have missed some eligible papers.

We agree with this comment and we have included this as a limitation of our study:

Response Line 28-29 (pg14)
"However, the use of only one database, PubMed, which is comprehensive but not exhaustive, may have missed eligible papers…"

• Additional criteria applied during the eligibility assessment has also excluded many papers.

While the strategy, restrictions and eligibility criteria are all sensible and there are clear reasons given for them, it is likely that the authors have overestimated the quality of reporting, possibly by quite some margin. Though this is acknowledged, I would like to see some discussion of the implications of this; if the academic community is taking the excluded 'pilot' trials seriously, understanding the quality of reporting therein is also important, arguably more so, since there are more of them and their prevalence may be contributing to continued poor reporting. In particular, the 32 trials excluded for not assessing feasibility would be of interest. Are they all simply underpowered main trials? If so, why might they be being legitimately published as pilot or feasibility studies? If not, what are they addressing, and to what end? A discussion of these points would be helpful to give a more accurate picture of the actual state of trials reported as pilot or feasibility trials.

We have included a paragraph within the manuscript to address this comment:

Response Line 46-54 (pg14) and 7-8 (pg15)
"During sifting, we identified 32 trials that had 'pilot' or 'feasibility' in the title/abstract, but were not assessing feasibility. A third of these were identified because they referred to 'pilot' or 'feasibility' at some point in the abstract but it was not in reference to the current trial (e.g. stating feasibility has already been shown), but the other two thirds were labelled as a pilot or feasibility trial yet showed no evidence of assessing feasibility and were only assessing effectiveness. This is an important point as our review may appear to overestimate reporting quality by not including these studies. That there are underpowered main trials being published as pilot or feasibility studies is something that the academic community should look to prevent."

Minor considerations include:
1. What is the 'written guidance' which was used by CC and CL for data extraction? (p5)

A document was produced which gave details on what to extract, to ensure that the independent data extractors were following the same rules during extraction. The information provided in the written

guidance document has been presented in the “further information” column of Appendix 2. We clarify this in the manuscript:

Response Line 21-22 (pg5)
“CC and CL independently extracted data from all papers selected for inclusion in the review, and followed rules on what to extract (see Further information column of Appendix 2).”

2. The items not included because ‘we expected the item would be generally well reported’ (p5) should be detailed in some way – at the least, report the number of items not included from each category, or if feasible, include these in an appendix so the reader can understand what was not assessed.

There was in fact only one item that fell in this category, which was given in brackets in the manuscript. We have reworded this sentence:

Response Line 18-19 (pg6)
“We also did not extract on whether papers reported a structured summary of trial design, methods, results, and conclusions.”

3. Similarly, how many/which items were excluded because they were difficult to extract? (p6)

Again, there were in fact only a few items that fell in this category, so we list them in the manuscript:

Response Line 36-41 (pg6)
“In particular: whether the objectives, intervention, or allocation concealment were at the individual level, cluster level, or both; and other analyses performed or other unintended consequences (difficult to decipher from papers whether it classified as an ‘other’).”

4. Incomplete reference 9 (p15)

Thank you for pointing this out. We have corrected this in the manuscript:

Response Line 28-30 (pg17)
“9. Díaz-Ordaz K, Kenward MG, Cohen A, Coleman CL, Eldridge S. Are missing data adequately handled in cluster randomised trials? A systematic review and guidelines. Clinical Trials. 2014;11(5):590-600.”

VERSION 2 – REVIEW

REVIEWER	Ruth Pickering University of Southampton UK
REVIEW RETURNED	28-Jun-2017

GENERAL COMMENTS	Much easier to read this version. I still found Table 4 difficult to follow. I didn't notice the key to the fonts used in the second column indicating where the items came from immediately. When reading the text describing results in Table 4 (page 43 onwards) was spending quite a bit of time identifying the items and results being commented upon. Otherwise a useful and well written paper.

REVIEWER	Jennifer Lewis

	ScHARR, University of Sheffield UK
REVIEW RETURNED	27-Jun-2017

GENERAL COMMENTS	Thank you for your clear and comprehensive responses to my original review. While I still have some reservations about the small number of papers included, I accept that this can be seen as representing a sample that indicates the current status of reporting, and that a less restrictive search is likely to have largely produced records that were more poorly reported, thus not altering the conclusions of the paper. I am satisfied that my concerns have been addressed, and am happy to recommend this paper for publication.

VERSION 2 – AUTHOR RESPONSE

Reviewer: 2
Reviewer Name: Jennifer Lewis
Institution and Country: ScHARR, University of Sheffield, UK Please state any competing interests or state 'None declared': None declared

Please leave your comments for the authors below Thank you for your clear and comprehensive responses to my original review. While I still have some reservations about the small number of papers included, I accept that this can be seen as representing a sample that indicates the current status of reporting, and that a less restrictive search is likely to have largely produced records that were more poorly reported, thus not altering the conclusions of the paper.

I am satisfied that my concerns have been addressed, and am happy to recommend this paper for publication.

Thank you.

Reviewer: 1
Reviewer Name: Ruth Pickering
Institution and Country: University of Southampton, UK Please state any competing interests or state 'None declared': NONE

Please leave your comments for the authors below Much easier to read this version. I still found Table 4 difficult to follow. I didn't notice the key to the fonts used in the second column indicating where the items came from immediately. When reading the text describing results in Table 4 (page 43 onwards) was spending quite a bit of time identifying the items and results being commented upon.
Otherwise a useful and well written paper.

We have added a comment about the table 4 footnote in the text of the manuscript to make Table 4 easier to follow.

Response Line 40-45 (pg42)
"The footnote of Table 4 also explains where the quality assessment items come from, with different font differentiating items based on the CONSORT extension for pilot trials and the CONSORT extension for CRTs, and a key to highlight which of the three categories above the item falls under."

To help identify where items being commented in the text are found in table 4, we have added a

couple of comments in the manuscript, and have explicitly stated the item numbers in brackets in one paragraph:

Response Line 55 (pg42)
“See items with [N] in column 2 of Table 4.”

Response Line 32 (pg43)
“See items with [S] in column 2 of Table 4.”

Response Line 17-36 (pg44)
“For the remaining items, reporting quality was variable. Some were reported by fewer than 20% of the pilot CRTs, for example considering the cluster design in the sample size rationale for the pilot trial (17%) (item 7a), reporting whether consent was sought from clusters (11%) and who enrolled them (17%) (items 10c and 10a), how people were blinded (7% of applicable trials) (item 11a), number of excluded individuals (6% of applicable trials) and clusters (18% of applicable trials) after randomisation (item 13b), and a table showing baseline cluster characteristics (11%) (item 15). Those reported most well, by more than 80% of the pilot CRTs, included reporting ‘pilot’ or ‘feasibility’ in the title (83%) (item 1a), scientific background and explanation of rationale for future definitive trial (100%) (item 2a), pilot trial design (100%) (item 3a), nature of the cluster (100%) (item 3a), settings and locations where the data were collected (100%) (item 4b), whether consent was sought from participants (94%) (item 10c), number of clusters randomised (94%) and assessed for primary objective (82% of applicable trials) (item 13a), number of individuals assessed for primary objective (94% of applicable trials) (item 13a), limitations of pilot trial (94%) (item 20), and source of funding (100%) (item 25).”

BMJ Open Editorial Office

1. Please remove Background on your Abstract section.

We have removed the background on our abstract.

Response Line 28-31 (pg35)
(deleted background)

2. Please embed your Competing interests statement on your main document file as shown in scholar one.

This is located on page 49 of the proof, lines 44-48, after the conclusion and contributors section. CC rang the BMJ Open associate publisher to confirm that the competing interest statement is embedded correctly.