PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Quality improvement needed in quality improvement randomized trials: Systematic review of interventions to improve care in diabetes
AUTHORS	Ivers, Noah; Tricco, Andrea; Taljaard, Monica; Halperin, Ilana; Turner, Lucy; Moher, David; Grimshaw, Jeremy

VERSION 1 - REVIEW

REVIEWER	Jonathan R. Treadwell, PhD, ECRI Institute, Associate Director of the Evidence-based Practice Center
REVIEW RETURNED	21-Feb-2013

GENERAL COMMENTS	Overall it's a well-done paper. My focus below is improvements/clarifications/suggested additions. Regarding your trend test. Lines 152-155 and Table 3. For individual domains, you investigate the trend of time in the *proportion of studies that had a low risk of bias for that domain* (because you combined high and unclear). When you later come to the analysis across domains, you analyze the *proportion of studies that had at least one high risk of bias domain.* The focus appears to change from Low to High, which is confusing. I advise something cleaner: Redo the one-domain-at-a-time analyses (the ones you say you didn't do because "the number of studies judged to have high risk of bias was very small for many individual domains") by **combining the Low and Unclear** categories. All tests are nonsignificant (I verified in stata for each of the 8 domains), so your overall message is the same (ie no evidence for a trend in quality). (The only test that came back 'undefined' is the last one because there are all 0 proportions of high risk of bias….that can be dealt with in a footnote in the new table 3 (test of this domain was undefined due to the absence of high risk of bias studies….therefore we did this one domain by instead combining the high and unclear categories and found no effect p=0.52). Also be sure to change the existing footnote to say that you combined low and unclear. Regarding present the trend test results only as p values. The problem is that p values are heavily influenced by the N, and so to only report them, without any effect size metrics or confidence intervals, has the potential to be misleading. What effect size metric corresponds to your Cochran-Armitage test? You should report that and its CI as the primary result, and suppress the p value, or else make the p value far less prominent.

	Same point about p values applies to Table 4, far right column. A measure of effect size would be more informative than a p value. Another question you could address with these data is whether *reporting* of risk-of-bias issues has improved over time. For this you would combine the high and low categories. I know, insanity, but still it does directly measure the trend in reporting. I did the 8 tests and none were statistically significant. So perhaps you want another message of your paper to be that reporting isn't improving either. Really all this takes is a few extra sentences…I don't see a need for a new table. Plus your abstract could say there is no evidence of recent change in quality OR reporting. Regarding the overall analysis. Yet another possibility would be to assess the trend in HOW MANY DOMAINS were high risk-of-bias. It could be that your measure of "At least one high risk domain" is not sensitive enough to trends. Line 49. Not clear how this 3rd bullet is either a strength or limitation of YOUR paper. Perhaps add a sentence saying "lack of reporting may reflect authors' beliefs that readers do not need that level of detail, rather than simple authors' ignorance". Or insert your own wording, if that's what you mean. Line 157. Add a citation for your asymptomatic CI (maybe Wilson score interval?) Line 180. Text says 4%, figure says 3%. It's 3.52%, 5/142, so the figure needs to be changed. Lines 248-256. This sensitivity test should have been described in the methods and mentioned in the results. Lines 244-266. This is a very long paragraph, try to chop it up into multiple paragraphs. Figure 1.There were 5 included RCTs that were in fact not RCTs (top of figure 1, dark segment indicating that 3.52% or 5/142 studies were not adequate sequence generation, which itself is defined as not randomized in table 1). Seems like those 5 should have failed the inclusion criteria.

VERSION 1 – AUTHOR RESPONSE

1) Regarding your trend test. Lines 152-155 and Table 3. For individual domains, you investigate the trend of time in the proportion of studies that had a low risk of bias for that domain (because you combined high and unclear). When you later come to the analysis across domains, you analyze the proportion of studies that had at least one high risk of bias domain. The focus appears to change from Low to High, which is confusing. I advise something cleaner: Redo the one-domain-at-a-time analyses (the ones you say you didn't do because "the number of studies judged to have high risk of bias was very small for many individual domains") by combining the Low and Unclear categories. All tests are nonsignificant (I verified in stata for each of the 8 domains), so your overall message is the same (ie no evidence for a trend in quality). (The only test that came back 'undefined' is the last one because there are all 0 proportions of high risk of bias….that can be dealt with in a footnote in the new table 3 (test of this domain was undefined due to the absence of high risk of bias studies….therefore we did this one domain by instead combining the high and unclear categories and found no effect p=0.52). Also be sure to change the existing footnote to say that you combined low and unclear.

Thank you for this suggestion. We agree that changing the focus from low to high is confusing. We have reanalyzed the data by combining the low and unclear categories. Because of the small numbers of studies with high risk of bias, we now present the exact (non-parametric) version of the Cochran-Armitage trend test. We have modified the footnote as suggested by the reviewer.

2) Regarding present the trend test results only as p values. The problem is that p values are heavily influenced by the N, and so to only report them, without any effect size metrics or confidence intervals, has the potential to be misleading. What effect size metric corresponds to your Cochran-Armitage test? You should report that and its CI as the primary result, and suppress the p value, or else make the p value far less prominent.Same point about p values applies to Table 4, far right column. A measure of effect size would be more informative than a p value.

We agree that effect sizes are more informative than p-values; however, we are not aware of effect size measures corresponding with the exact Cochran-Armitage trend test. Pairwise differences between the years could be presented, but there would be 3 confidence intervals associated with each domain, and we do not think that these would be helpful. Moreover, the proportions of studies with high risk of bias are very small. We feel that presenting the proportions in the 3 categories of publication year is adequate to allow the reader to judge the absence of any improvement over time. Likewise, Table 4 presents an exploratory analysis of factors potentially associated with at least one high risk of bias domain. These are overall chi-squared tests of proportions, and we feel that adding multiple odds ratios and confidence intervals would unnecessarily complicate the table.

3) Another question you could address with these data is whether reporting of risk-of-bias issues has improved over time. For this you would combine the high and low categories. I know, insanity, but still it does directly measure the trend in reporting. I did the 8 tests and none were statistically significant. So perhaps you want another message of your paper to be that reporting isn't improving either. Really all this takes is a few extra sentences…I don't see a need for a new table. Plus your abstract could say there is no evidence of recent change in quality OR reporting.

This is an interesting suggestion, thank you. We have calculated the trend tests for reporting as the reviewer suggested and have added this to the objectives, methods, and results.

4) Regarding the overall analysis. Yet another possibility would be to assess the trend in HOW MANY DOMAINS were high risk-of-bias. It could be that your measure of "At least one high risk domain" is not sensitive enough to trends.

Thanks you again - this has some face validity, but our argument for this paper was that if even one

domain is at high risk, the entire result must be viewed with a grain of salt – in that sense it matters little whether a trial has one or four domains at high risk of bias. Nevertheless, we looked at the numbers of domains with high risk over time. Analysis is complicated by the fact that only half of studies had a non-zero count; and among those, most had only one domain with high risk. Exact counts are 73 had 0 domains, 44 had one, 19 had two, 5 had three and one had four. The mean counts over time were 0.59, 0.61, and 0.86, which is difficult to interpret given the distribution of the counts. The p-value for a difference over time was 0.22. We would prefer to not include this analysis in the manuscript, as we believe it would risk over-complicating the message.

5) Line 49. Not clear how this 3rd bullet is either a strength or limitation of YOUR paper. Perhaps add a sentence saying “lack of reporting may reflect authors’ beliefs that readers do not need that level of detail, rather than simple authors’ ignorance”. Or insert your own wording, if that’s what you mean.

This bullet has been altered to clarify the meaning. It now more clearly acknowledges that risk of bias is not a comprehensive approach to assigning value to a trial.

6) Line 157. Add a citation for your asymptomatic CI (maybe Wilson score interval?)

These are standard Wald asymptotic confidence limits for the binomial proportion. We do not have a specific reference for this.

7) Line 180. Text says 4%, figure says 3%. It’s 3.52%, 5/142, so the figure needs to be changed.

Thank you for catching this typo – the figure has been updated.

8) Lines 248-256. This sensitivity test should have been described in the methods and mentioned in the results.

Thank you for this suggestion. The manuscript has been revised accordingly.

9) Lines 244-266. This is a very long paragraph, try to chop it up into multiple paragraphs.

Thank you for this suggestion. The manuscript has been revised accordingly.

10) Figure 1.There were 5 included RCTs that were in fact not RCTs (top of figure 1, dark segment indicating that 3.52% or 5/142 studies were not adequate sequence generation, which itself is defined as not randomized in table 1). Seems like those 5 should have failed the inclusion criteria.

This is an interesting observation. Many studies describe themselves as RCTs but in fact would be better described as CCT. While we agree in principle that studies not ‘truly’ randomized should not be considered RCTs, this was a secondary analysis of an established systematic review. Our careful assessment of risk of bias according to the criteria described in the manuscript allowed us to identify these issues.