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)	Effect of digital health, biomarker feedback and nurse or midwife-led counselling interventions to assist pregnant smokers quit: A systematic review and meta-analysis
AUTHORS	Tahan, Chadi; Dobbins, Timothy; Hyslop, Fran; Lingam, Raghu; Richmond, Robyn

VERSION 1 – REVIEW

REVIEWER	Løkkegaard, Ellen Nordsjaellands Hospital, Department of Gynecology and Obstetrics
REVIEW RETURNED	21-Mar-2022

GENERAL COMMENTS	This is a systematic review and meta-analysis of randomized controlled trials on communication technology, biomarker feedback and nurse or midwife counselling to assist pregnant smokers to quit smoking. The review is performed with PRISMA, PICO and GRADE and seem very well conducted. My only concern is the abbreviations I keep forgetting. Minor comment Consistent use of capital letters in tables and 1000 separators.

REVIEWER	Patnode, CD Kaiser Permanente Medical Center
REVIEW RETURNED	10-May-2022

GENERAL COMMENTS	This is a fine review, with clear methods. My biggest concern is regarding the rationale for the review, and in particular, the need and focus on these 3 specific types of interventions. There is a fairly recent Cochrane review (Chamberlain 2017) and update for the USPSTF (Patnode 2021) that synthesizes this evidence. Furthermore, the article is extremely long. In revising, I would suggest including higher-level summaries of your findings and combining narrative and meta-analysis results. There is considerable trimming that can be done. You might consider a Table 1 that presents your meta-analytic results for all 3 types of interventions, for CA and PPA, at all time points. Then, you can make high-level summaries of where there were and were not significant associations in meta-analysis. The narrative results should support your meta-analysis results or vice versa. Other secondary notes: - “Communication technologies” may not be the preferred terminology to use to describe these types of interventions. You

	may want to consider “Technology-based interventions,” “Mobile interventions,” “Digital interventions,” or “Telehealth-based interventions” as alternatives. -I would include the narrative results of CT and BF in the main text, and cut words elsewhere. The narrative results should still be at a high-level, and not necessarily results for all individual studies. - Similarly, in your quality assessment section, I would present a high-level summary of the major domains of risk of bias versus getting into individual study ratings. - In your results of the meta-analysis, I would suggest include the number of studies that were pooled in each analyses (e.g., RR = 1.37, CI = , # studies). - Consider making an overall statement of the findings as the first sentence within each intervention category. For example, “In pooled analyses, there was a statistically significant association between communication technology interventions and point prevalence abstinence in the postpartum period. In late pregnancy, there was no association between these interventions and point prevalence abstinence but there was for measures of continuous abstinence.” Also, this is an odd finding, and the discussion should address this! -I would consider making Table 1 a summary of the effect sizes seen for each type of intervention and outcome, including one measure of heterogeneity rather than all measures of heterogeneity (and no effects). - The “P” in PICO stands for “Population” -Page 9, line 27 “risk of bias for each “study” (not intervention) -There is duplication on page 9 lines 53-63 (in regards to starting RCTs as having high certainty) - Please include information on how you assessed the risk of bias for non-RCT studies.

REVIEWER	Rogers, Kris University of Technology Sydney
REVIEW RETURNED	03-Aug-2022

GENERAL COMMENTS	Thanks for the opportunity to review your manuscript. My role is as a statistical reviewer, so my review concentrates on the study design, data, and analysis that are presented. I have put general questions first, followed by queries relevant to a specific section of the manuscript (with a page/line reference). This manuscript presents a systematic review and meta-analysis of nurse/midwife counselling, bio-marker feedback, and communication technology interventions intended to lead to smoking cessation in pregnant women. The three types of interventions were considered separately (appropriately, in my non-content-matter-expert opinion). Usual care, wait-list, and other comparators not part of the main group of interventions were considered as comparators to the selected interventions. Several ways of measuring abstinence (point, continuous) with and without biochemical validation are considered. P-curves were the key check of publication/pre-publication bias in the included studies. The review is comprehensive, the detailed material in the supplementary appendix was helpful in completing the review, thank you. Overall the statistical approach is this manuscript is good. The random-effects model used is fine and the p-curve approach to assessing pre/post-publication bias is appropriate. I agree with the

	main interpretation of the p-curves of not enough sample to assess bias for CT (the p-curve has very few studies <0.05 to make an assessment about flatness or skewness), and no evidence of bias for the other types of interventions. Given the other methods (e.g. funnel plot, Egger) are presented with conflicting results to the p-curve, how should these be interpreted? Are the 'significant' test results a consequence of the number of studies or does it reflect a serious issue? Was the protocol for this review registered (e.g. PROSPERO)? P6, L36. I would reword this given that strictly speaking, guidelines like PRISMA specify how to report a study rather than prescribing a specific methodology to be used. P10, L25. Which specific effect size measure was used when converting? P17, L41. I couldn't manage to find what type of outcome these three excluded studies had that mean it would be unable to be converted to an effect size measure. P18, L19. I'd just rephrase this – as a moderate level of heterogeneity isn't evidence in itself in favour/not in favour of interventions.

VERSION 1 – AUTHOR RESPONSE

Reviewer: 1
Prof. Ellen Løkkegaard, Nordsjaellands Hospital
Comments to the Author:
This is a systematic review and meta-analysis of randomized controlled trials on communication technology, biomarker feedback and nurse or midwife counselling to assist pregnant smokers to quit smoking.
The review is performed with PRISMA, PICO and GRADE and seem very well conducted.
My only concern is the abbreviations I keep forgetting.
Minor comment

4. Consistent use of capital letters in tables and 1000 separators.

We thank the reviewer for highlighting these inconsistencies and the proposal to use separators. We have applied changes in all tables to address these issues.

Reviewer: 2
Dr. CD Patnode, Kaiser Permanente Medical Center
Comments to the Author:

5. This is a fine review, with clear methods. My biggest concern is regarding the rationale for the review, and in particular, the need and focus on these 3 specific types of interventions. There is a fairly recent Cochrane review (Chamberlain 2017) and update for the USPSTF (Patnode 2021) that synthesizes this evidence.

We thank the reviewer for this important observation. We agree that other research works have investigated and presented evidence on effectiveness of various interventions. However, this review focused on three specific inventions that were being considered together for potential inclusion in a

clinical trial in relevant clinical settings for pregnant smokers. The review enabled consistent approach to evaluating the three interventions.

6. Furthermore, the article is extremely long. In revising, I would suggest including higher-level summaries of your findings and combining narrative and meta-analysis results.

There is considerable trimming that can be done. You might consider a Table 1 that presents your meta-analytic results for all 3 types of interventions, for CA and PPA, at all time points. Then, you can make high-level summaries of where there were and were not significant associations in meta-analysis. The narrative results should support your meta-analysis results or vice versa.

We thank the reviewer for his suggestion. We have followed his suggestion and constructed Table 1 with summary of all effect estimates. We have also summarised the narrative synthesis and updated this result section as advised.

Other secondary notes:
7. “Communication technologies” may not be the preferred terminology to use to describe these types of interventions. You may want to consider “Technology-based interventions,” “Mobile interventions,” “Digital interventions,” or “Telehealth-based interventions” as alternatives.

We thank the reviewer for his suggestion and we have changed the description of this intervention to digital health throughout the manuscript and in all supplementary files.

8. I would include the narrative results of CT and BF in the main text, and cut words elsewhere. The narrative results should still be at a high-level, and not necessarily results for all individual studies.

We thank the reviewer for this recommendation. We have now incorporated the narrative results for DH and BF in the main text and summarised the narrative results for NoMC interventions.

9. Similarly, in your quality assessment section, I would present a high-level summary of the major domains of risk of bias versus getting into individual study ratings.

We thank the reviewer for his suggestion and we have edited the risk of bias results section in an attempt to focus on major domains for the bias rather than focusing on individual studies.

10. In your results of the meta-analysis, I would suggest include the number of studies that were pooled in each analyses (e.g., RR = 1.37, CI = , # studies).

We thank the reviewer for his suggestion. Although we have included the number of studies that were pooled in the Summary of Evidence table for each intervention, we have now also included these in the results section for each intervention.

11. Consider making an overall statement of the findings as the first sentence within each intervention category. For example, “In pooled analyses, there was a statistically significant association between communication technology interventions and point prevalence abstinence in the postpartum period. In late pregnancy, there was no association between these interventions and point prevalence abstinence but there was for measures of continuous abstinence.” Also, this is an odd finding, and the discussion should address this!

We thank the reviewer for his suggestion. We have followed his suggestion and included summative statements that describe the various overall findings for each intervention category.

12. I would consider making Table 1 a summary of the effect sizes seen for each type of intervention and outcome, including one measure of heterogeneity rather than all measures of heterogeneity (and no effects).

We thank the reviewer for his recommendation and acknowledge this table 1 would be important to provide a summary of overall findings across all three interventions. We have now included Table 1 summary of effect size table under 'Statistical analysis' in the Results section of the manuscript.

13. The "P" in PICO stands for "Population"

We thank the reviewer for his correction. Pregnant smokers should not be regarded as patients and pregnancy is not a disease. We have updated the manuscript to ensure that P stands for Population not patient.

14. Page 9, line 27 "risk of bias for each "study" (not intervention)

We thank the reviewer for highlighting this issue. We have edited the text for better clarity.

15. There is duplication on page 9 lines 53-63 (in regards to starting RCTs as having high certainty)

We thank the reviewer for picking up this duplication. We have now removed the duplication.

16. Please include information on how you assessed the risk of bias for non-RCT studies.

We thank the reviewer for this suggestion. A brief explanation of the approach taken to assess risk of bias, including for non-RCT studies has now been added under 'Risk of bias assessment and certainty of evidence' subheading in the Method section.

Reviewer: 3
Dr. Kris Rogers, University of Technology Sydney
Comments to the Author:
Thanks for the opportunity to review your manuscript. My role is as a statistical reviewer, so my review concentrates on the study design, data, and analysis that are presented. I have put general questions first, followed by queries relevant to a specific section of the manuscript (with a page/line reference). This manuscript presents a systematic review and meta-analysis of nurse/midwife counselling, bio-marker feedback, and communication technology interventions intended to lead to smoking cessation in pregnant women. The three types of interventions were considered separately (appropriately, in my non-content-matter-expert opinion). Usual care, wait-list, and other comparators not part of the main group of interventions were considered as comparators to the selected interventions. Several ways of measuring abstinence (point, continuous) with and without biochemical validation are considered. P-curves were the key check of publication/pre-publication bias in the included studies. The review is comprehensive, the detailed material in the supplementary appendix was helpful in completing the review, thank you.

Overall the statistical approach is this manuscript is good. The random-effects model used is fine and the p-curve approach to assessing pre/post-publication bias is appropriate. I agree with the main interpretation of the p-curves of not enough sample to assess bias for CT (the p-curve has very few studies <0.05 to make an assessment about flatness or skewness), and no evidence of bias for the other types of interventions.

17. Given the other methods (e.g. funnel plot, Egger) are presented with conflicting results to the p-curve, how should these be interpreted? Are the 'significant' test results a consequence of the number of studies or does it reflect a serious issue?

The conflicting results indicated that the application of more traditional small study tests such as Funnel Plot and Egger's test may not lead to a comprehensive assessment of publication bias. More importantly, unlike the more conventional tests, the p-curve test allows us to determine if the data contain an effect that exists in reality (i.e. the effect is not simply spurious and exists only due to selective reporting).

The Funnel Plot, Egger's test and Duval and Tweedie Trim and Fill methods attempt to assess and correct for small study bias (which may include publication but also other types of biases such as research design). These approaches are based on the assumption that publication bias is disproportionately affected by small studies and they are therefore somehow limited as they may not truly test for publication bias in a comprehensive or a more representative way. This is because these approaches do not address the problem of p-hacking.

The P-curve test on the other hand can be considered a relatively novel and a more accurate approach to assessing publication bias. This is because the p-curve assesses publication bias not based on effect size but rather a plot of p values. This allows us to examine p-hacking, which is what is more likely to happen in practice. Therefore, we believe that the use of p-curse as a test of publication bias may be more representative of assessing for publication bias than other traditional approaches that are based on small study bias.

Funnel plot allows a quick view of effect sizes of all studies to establish if publication bias is likely through a visual inspection of symmetry in the plot. This can be subjective and thus the Egger's test is an objective alternative to verify the Funnel Plot results.

It is advisable to test for possible publication bias through funnel plot asymmetry inspection only when there is sufficient number of studies to provide enough statistical power. This is because low number of studies may not provide statistical power high enough to indicate asymmetry. It can be argued that this would be the case for DH interventions.

Further, the Trim and Fill Duval does not produce reliable results when between study heterogeneity is large/considerate. Also, the Trim and Fill method only attempts to estimate the corrected effect size but does not correct that effect size accurately. Therefore, when large studies do not share one true effect, then applying trim and fill method may result in inaccurate results.

Finally, as for the p-curve test, a 'no-no' outcome as witnessed from the DH interventions indicate that we cannot confirm that a true effect exists but we can not equally rule out that a very small effect exists either. This outcome is indicates that more studies will result in a different p-curve test outcome.

18. Was the protocol for this review registered (e.g. PROSPERO)?

We would like to confirm that this review did not have prior registration, including in PROSPERO.

19. P6, L36. I would reword this given that strictly speaking, guidelines like PRISMA specify how to report a study rather than prescribing a specific methodology to be used.

We thank the reviewer for pointing this out. We have now adjusted the wording to clarify the approach used.

20. P10, L25. Which specific effect size measure was used when converting?

The effect sizes we used were the whole number of individuals (rounded to 1) that had been reported and biochemically confirmed to have achieved abstinence (PPA or CA) at each time point for each arm of a study. We sourced those numbers from the published articles of studies. These numbers are shown in Supplementary file 3

21. P17, L41. I couldn't manage to find what type of outcome these three excluded studies had that mean it would be unable to be converted to an effect size measure.

The three excluded studies are as shown below and yes, measures could not be confidently converted to effect size measures that would be comparable to other studies:
Study Outcome
1. Tombor et al 2019; Smoke free days 4 weeks after quit date
2. Zhang et al 2017; Abstinence during or post intervention of high vs low attendance (low attendance outcome not reported)
3. Joseph et al 2009; PPA at second and third trimester follow up reported as risk of Smoking Odds Ratio.

22. P18, L19. I'd just rephrase this – as a moderate level of heterogeneity isn't evidence in itself in favour/not in favour of interventions.

We thank the reviewer for his comment. We acknowledge the previous description of the results was not very clear. Following other recommendations, the description of results in this section has now been updated to provide better clarity.

Reviewer: 1
Competing interests of Reviewer: No competing interests

Reviewer: 2
Competing interests of Reviewer: None

Reviewer: 3
Competing interests of Reviewer: I was previously employed in the same university department as some of the authors. I have not worked or collaborated directly with any of the authors.

VERSION 2 – REVIEW

REVIEWER	Patnode, CD Kaiser Permanente Medical Center
REVIEW RETURNED	10-Nov-2022

GENERAL COMMENTS	Thank you for your responses and revisions to the manuscript. It is improved and could be considered for publication. -I remain concerned about the length of the article and feel there is significant text that could be cut or tightened. If information can be found in tables, you may not need the level of detail that is currently presented narratively (for example, the section on comparators of DH studies could simply read: Comparators for DH interventions varied across studies including usual care or brief advice (80, 81, 83, 84, 89, 98), an attention control (86, 87, 134

	add 97 here?), self-help materials (90, 93, 96, 135) and an active comparator of telephone counseling (94). -The article would also greatly benefit from line-editing. -I remained concerned about how the ROB was assessed for non-randomized studies. Are you saying you used the Cochrane RoB tool (which is designed for randomized studies) for all designs? Perhaps comment on why you didn't use a tool that is designs for NRSIs (like ROBINS-I) -You could consider cutting the details about the levels of GRADE ratings. That whole paragraph could be trimmed to "We used GRADE to assess the overall certainty of evidence taking into account the risk of bias of individual studies, consistency, directness, precision, and publication bias."

REVIEWER	Rogers, Kris University of Technology Sydney
REVIEW RETURNED	11-Nov-2022

GENERAL COMMENTS	Thanks for the revised manuscript and responses to my review. I appreciate the explanation about the application of the p-curve method –I think that the methods applied to assess publication bias and p-hacking are appropriate and have been interpreted correctly. The revisions to the manuscript have cleared up my original questions and I recommend that the manuscript be accepted.

VERSION 2 – AUTHOR RESPONSE

Reviewer: 2. Dr. CD Patnode, Kaiser Permanente Medical Center Comments to the Author:

1. Thank you for your responses and revisions to the manuscript. It is improved and could be considered for publication.

Thank you for your feedback.

2. I remain concerned about the length of the article and feel there is significant text that could be cut or tightened. If information can be found in tables, you may not need the level of detail that is currently presented narratively (for example, the section on comparators of DH studies could simply read: Comparators for DH interventions varied across studies including usual care or brief advice (80, 81, 83, 84, 89, 98), an attention control (86, 87, 134 add 97 here?), self-help materials (90, 93, 96, 135) and an active comparator of telephone counseling (94).

Thank you for this feedback. We have made further attempts to reduce the length of this manuscript.

The article would also greatly benefit from line-editing.

Thank you for this feedback. We have applied line-editing to help improve the manuscript's readability.

3. I remained concerned about how the ROB was assessed for non-randomized studies. Are you saying you used the Cochrane RoB tool (which is designed for randomized studies) for all designs? Perhaps comment on why you didn't use a tool that is designs for NRSIs (like ROBINS-I)

Thank you for this valid concern. As previously stated, we have applied the RoB-2 tool to assess bias across all included studies. The reason for this approach was to elicit a consistent approach in assessing this type of potential bias.

Most of the included studies were RCTs, and as you have correctly highlighted, some were not. We did take into consideration the additional potential sources of bias, namely confounding, selection and information bias, as much as possible when using the RoB-2 tool.

We do acknowledge however that using the ROBINS-I tool to assess the intra-study bias for these (non-RCT) studies would have been more appropriate and may have produced more accurate assessment results.

However, we believe that in this particular case (i.e. the use of ROBINS-I for these studies alongside RoB-2 tool for RCTs) would not lead to any material change in the reported overall results and conclusions. This is because there were 14 non RCT studies that were included in this review. None of these studies were deemed to have low risk of bias, only 4 has been rated to have some concerns and most (i.e. 10) were deemed to have high risk of bias, using the RoB-2 tool.

All 4 studies with some concerns in relation to intra study bias were related to the nurse or midwife counselling intervention type. Two of these four studies had PPA measures in late pregnancy (Everett_Murphy et al 2010 and Wisborg et al 1998), De Vries et al 2006 had PPA in postpartum and Gebauer et al 1998 had PPA at follow up but this measure was not included in the meta-analysis.

Non-RCT Study Study Type RoB Assessment result

Forinash et al 2018 Randomised prospective intervention study High risk

Windsor et al 2011 Quasi experimental evaluation trial High risk

Zhang et al 2017 Real life controlled trial High risk

Wesselink et al 2015 Quasi experimental trial High risk

Hayes et al 2013 Quasi experimental trial High risk

Everett-Murphy et al 2010 Quasi experimental trial Some concerns

Edwards et al 2009 Prospective intervention study High risk

Oien et al 2008 Real life controlled trial High risk

De Vries et al 2006 Quasi experimental trial Some concerns

Britton et al 2006 Quasi experimental trial High risk

Ferreira-Borges et al 2005 Quasi experimental trial High risk

Hegaard et al 2003 Quasi experimental trial High risk

Wisborg et al 1998 Prospective intervention study Some concerns

Gebauer et al 1998 Quasi experimental trial Some concerns

4. -You could consider cutting the details about the levels of GRADE ratings. That whole paragraph could be trimmed to "We used GRADE to assess the overall certainty of evidence taking into account the risk of bias of individual studies, consistency, directness, precision, and publication bias."

Thank you for this feedback. We agree with your suggestion and have implemented it to reduce the length of this manuscript.

5. Reviewer: 3 - Dr. Kris Rogers, University of Technology Sydney Comments to the Author:

Thanks for the revised manuscript and responses to my review.

I appreciate the explanation about the application of the p-curve method –I think that the methods applied to assess publication bias and p-hacking are appropriate and have been interpreted correctly.

The revisions to the manuscript have cleared up my original questions and I recommend that the manuscript be accepted.

Thank you for your feedback.