PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	What components of smoking cessation care during pregnancy are implemented by health providers? a systematic review and meta-analysis
AUTHORS	Gould, Gillian; Twyman, Laura; Stevenson, Leah; Gribbin, Gabrielle; Bonevski, Billie; Palazzi, Kerrin; Bar Zeev, Yael

VERSION 1 - REVIEW

REVIEWER	Felix Naughton University of East Anglia, UK
REVIEW RETURNED	14-Sep-2018

GENERAL COMMENTS	This is a challenging literature and the authors have made substantial efforts to summarise it meaningfully. This has taken skill and it is well written. I have a few concerns that are probably quite challenging to address and a few suggestions for improvements. Most of my comments are orientated around heterogeneity. I was curious why the authors did not consider a priori looking at study design and respondent (HP vs. pregnant women) as potential sources of heterogeneity? After the studies were screened, I can see that most were surveys but it might not have been clear from the outset, but one would imagine discrepancies between what individual pregnant women reported in terms of the 5 As 'received' and rough estimates from HPs regarding whether they applied the 5 As often/always etc. The respondant aspect could be investigated in an exploratory analysis? I could imagine some considering these populations to be too different to combine into a meta-analysis, but perhaps if this wasn't a source of heterogeneity this would reassure the reader that it might be acceptable to do this? I am not sure I fully understood the difference between 'yes' and 'often/always'? How do we interpret what this means and which would we imagine might be closer to 100% of women asked, assisted etc.? The data seemed fairly different between these two categories. Also, how do HPs interpret 'often'? Would this be 50% of the time, 80% of the time? Was this operationalised in any of the included studies? Perhaps the authors could comment on this in the discussion (apologies if I missed this). It makes it hard to interpret what this means e.g. 90% of HPs ask pregnant women

	between 50-80% of the time. Greater discussion/consideration of this would strengthen the paper in my view. I wasn't clear if for 'Ask' studies were providing a % of HPs to describe whether they asked non-pregnant smokers about whether they smoked or not? And then for 'Assess', was this based only on pregnant smokers i.e. those who said 'yes' to Ask? These are quite different populations, so it would be good to have clarity throughout paper on this. Some of the meta-analysis findings don't make intuitive sense. For example, fewer HPs reported often or always assisting with unspecified cessation support (59%) than HPs who said 'Yes' they assisted with a counselling offer (81%), which is a specific treatment. Understandably these differences are likely to be because they pool different studies, but this reflects the high heterogeneity. Not much you can do about this other than acknowledge the issues comparing different operationalisations of the 5 As. But it reduces the reader's confidence in the appropriateness of comparing these head to head. I wasn't sure why heterogeneity was not assessed for meta-analyses assessing less than 5 studies using a fixed effects model. Perhaps more importantly, I was also not sure I understood the reason for undertaking a fixed effects analysis for area with equal to or less than 5 studies. Perhaps the justification provided can be supported by a reference to help educate the readers, as I had not come across this practice before? I don't understand why power has anything to do with this, given there was no hypothesised proportion the review was aiming to identify. Presumably it simply just effects precision in terms of the 95% CI width of the pooled estimate (which is usually larger in random effects MA anyway). Given the high heterogeneity, it doesn't seem at face value appropriate to undertake fixed effects meta-analysis of studies with many varying characteristics given this approach makes an assumption that there is one 'true' proportion? This was done for ~10 out of 17 of the meta-analyses, so it is not trivial. I wondered if effect sizes/statistical results should be included in second paragraph (starting 'table 1…') on page 16? It was odd for table 1 to be the meta-regression findings. I would have expected at least a summary of results as the first table rather than an exploration of heterogeneity. Figure 2 should come before table 1 at least? Also, it is not immediately clear why models were not generated for some combinations in table 1? Table 3 – not immediately clear that this refers to quality scoring. Suggest tweak to title

REVIEWER	Lucinda England Centers for Disease Control and Prevention USA
REVIEW RETURNED	17-Oct-2018

GENERAL COMMENTS	Thank you for the opportunity to review What components of smoking cessation care during pregnancy are implemented by health providers? A systematic review and meta-analysis. In this

	study, the authors collected studies of administration of components of the 5 As by providers to pregnant women and calculated pooled estimates of the percentage of providers performing each of the 5 As. General comments: My main concern about this approach is whether there is any useful information to be gained by pooling data for this type of estimate. The logical follow up if administration of the 5 As is low would be to implement interventions to increase use of the 5 As. However, by pooling data, you are masking findings that would be useful for follow up, hiding the places where administration is low. The same is true for including years that aren't recent. I'm concerned that data that are decades old don't reflect current practice and so shouldn't be combined with recent data. Alternative approaches would be to restrict to recent years, examine countries or localities separately, and/or focusing on particular areas rather than all countries. Specific comments Abstract I found the non-standard abbreviations (such as HP and SCC) to be distracting rather than helpful. I would spell these out. Introduction Page 5: there are more recent and more comprehensive resources documenting the effects of smoking on pregnancy outcomes. Suggest referencing the most recent Surgeon General's reports and review articles. Some of the outcomes associated with prenatal smoking (like childhood cancers) are not established. Line 40-41: less likely to abstain than whom? It isn't clear why the authors decided to look at NRT use in pregnant women in this analysis, given that many countries do not include a recommendation to use NRT in their clinical guidelines. Which countries recommend the 5 As vs. some other type of behavioral intervention? Methods Suggest restricting to recent studies (see general comments). Suggest restricting analysis of NRT to countries or localities where it is recommended for pregnant women. Page 8: How were studies of knowledge and attitudes and "other BCTs" incorporated into the analysis? The methods section here needs to be expanded. Results Page 12: some of the outcomes come up unexpectedly. Is there any evidence base for measuring exhaled CO as a cessation strategy?

	Page 16: the description of the authors' assessment of study quality (most had some aspects rated as good) does not give the reader an appreciation for the overall study quality of this body of literature. For example, what types of limitations were commonly found? How many studies were of high quality overall or in key areas? Inter-rater agreement wasn't optimal. What did the authors do to reconcile this? Discussion Page 20: There is not a strong evidence base for pharmacotherapy for pregnant women, with without psychosocial support. Therefore, it may not be bad or surprising that providers aren't prescribing NRT.

REVIEWER	Eirini Karyotaki Vrije Universiteit of Amsterdam
REVIEW RETURNED	06-Mar-2019

GENERAL COMMENTS	The paper describes the results of a systematic review and meta-analysis on the pooled prevalence rates for health providers in providing various components of smoking cessation care to pregnant women. Since the topic is outside of my expertise, I specifically looked at the use of the meta-analytic methodology. In general, the authors have performed this systematic review thoroughly and have provided a comprehensive overview of the current state of the art in this field. The meta-analytic methods are properly conducted. My main concern about this paper is the very high heterogeneity observed in the pooled prevalence rates (e.g., 99.1%). This very high heterogeneity represents a great variability among the rates reported by the examined studies, suggesting that the use of meta-analysis is questionable. Although the high heterogeneity is a very common issue in meta-analyses of prevalence rates, it is still not clear to me whether performing such meta-analyses is justifiable. The authors should consider the possibility of not pooling the rates of these very heterogeneous studies and focus on the narrative synthesis of the results. At the very least, the authors should try to explore what factors influence the variability of the prevalence rates and discuss the observed heterogeneity in details.

VERSION 1 – AUTHOR RESPONSE

Reviewer: 1

2. Most of my comments are orientated around heterogeneity. I was curious why the authors did not consider a priori looking at study design and respondent (HP vs. pregnant women) as potential sources of heterogeneity? After the studies were screened, I can see that most were surveys but it might not have been clear from the outset, but one would imagine discrepancies between what individual pregnant women reported in terms of the 5 As 'received' and rough estimates from HPs regarding whether they applied the 5 As often/always etc. The respondant aspect could be investigated in an exploratory analysis? I could imagine some considering these populations to be too

different to combine into a meta-analysis, but perhaps if this wasn't a source of heterogeneity this would reassure the reader that it might be acceptable to do this?

Response: To clarify: papers describing women's reports were analysed separately from those describing health provider reports.

This text added on P10. "Papers describing women's reports were analysed separately from those describing health provider reports."

We have added a comment to 'limitations' regarding the lack of comparison – we had considered that there may be differences in reporting the source of the data (HP vs patients) but considered that the populations would be too different to combine both for characteristics and clinical meaning. If fact, there was only one measure that was feasible to combine and when we tried to combine since receiving the review, this did not improve heterogeneity.

3. I am not sure I fully understood the difference between 'yes' and 'often/always'? How do we interpret what this means and which would we imagine might be closer to 100% of women asked, assisted etc.? The data seemed fairly different between these two categories. Also, how do HPs interpret 'often'? Would this be 50% of the time, 80% of the time? Was this operationalised in any of the included studies? Perhaps the authors could comment on this in the discussion (apologies if I missed this). It makes it hard to interpret what this means e.g. 90% of HPs ask pregnant women between 50-80% of the time. Greater discussion/consideration of this would strengthen the paper in my view.

Response: A few papers did define often and always as a percentage range, but not all, e.g. Bar-Zeev's paper defined always as 75% of the time or more, and often as 50-74%. 'Often' and 'always' response options were usually available as a Likert Scale then combined later for analysis as 'often/always' by the papers. Conceptually by using a scale to quantify responses are different from a 'yes /no'– which may be an option chosen by respondent whether they perform the practice anywhere from occasionally to frequently (ie not at all quantified) – therefore we did not combine often/always with yes/no study measures. In the discussion we have added:

P17: "Conceptually, using a scale to quantify responses is quite different from a 'yes' option: the latter may be an option chosen by respondent whether they perform the practice at an frequency from occasionally to always (ie not at all quantified) – therefore we did not combine 'often/always' with 'Yes/No' study measures."

Additionally to further clarify how outcome measures were combined we have added a supplementary text file as mentioned below:

P9: “General principles applied were as followed (explained in more detail in Supplementary Text 1):”

4. I wasn’t clear if for ‘Ask’ studies were providing a % of HPs to describe whether they asked non-pregnant smokers about whether they smoked or not? And then for ‘Assess’, was this based only on pregnant smokers i.e. those who said ‘yes’ to Ask? These are quite different populations, so it would be good to have clarity throughout paper on this.

Response: After a filter question for ‘Ask’ the other categories in most part (where described) applied to women only if they did smoke (e.g. out of 31 papers reporting ‘Advise’, 26 of them stated it was to women who smoked, and the rest were unclear. We thus reported on the measures as presented in the papers irrespective of whether a filter question was used.

5. Some of the meta-analysis findings don’t make intuitive sense. For example, fewer HPs reported often or always assisting with unspecified cessation support (59%) than HPs who said ‘Yes’ they assisted with a counselling offer (81%), which is a specific treatment. Understandably these differences are likely to be because they pool different studies, but this reflects the high heterogeneity. Not much you can do about this other than acknowledge the issues comparing different operationalisations of the 5 As. But it reduces the reader’s confidence in the appropriateness of comparing these head to head.

Response: Agreed – saying yes to ‘counselling’ does not indicate the frequency of performance – as we made in point #3. This is reflective of the different ways studies asked these questions, and we make a stronger comment in the discussion:

P17: “We acknowledge that there was no ideal way to combine these measures.”

6. I wasn’t sure why heterogeneity was not assessed for meta-analyses assessing less than 5 studies using a fixed effects model. Perhaps more importantly, I was also not sure I understood the reason for undertaking a fixed effects analysis for area with equal to or less than 5 studies. Perhaps the justification provided can be supported by a reference to help educate the readers, as I had not come across this practice before? I don’t understand why power has anything to do with this, given there was no hypothesised proportion the review was aiming to identify. Presumably it simply just effects precision in terms of the 95% CI width of the pooled estimate (which is usually larger in random effects MA anyway). Given the high heterogeneity, it doesn’t seem at face value appropriate to undertake fixed effects meta-analysis of studies with many varying characteristics given this approach makes an assumption that there is one ‘true’ proportion? This was done for ~10 out of 17 of the meta-analyses, so it is not trivial.

Response: Heterogeneity was assessed for the fixed effects modelling with small numbers of studies however in all cases it was found to be 0%; we believe this is not a true representation of the heterogeneity or inconsistencies of the study estimates, but more a product of the small number of studies and have chosen not to present these I-squared values. In this particular meta-analysis, it is correct that we are not interested in power, and we have changed the methods as follows:

P10:” If the number of studies was low (≤5), fixed effects modelling was used as the between-studies variance (tau-squared), and therefore the mean of the underlying random distribution cannot be estimated with precision; heterogeneity is not presented.”

References for the concern in undertaking a random-effects meta-analysis have been added, and while there may be more complex analytical methods available to estimate the average effect size, we have chosen to report and cautiously interpret the fixed effect method, acknowledging the limitation of this. We have added on P17 a caution about interpretation:

P 17 “resulting in small numbers of studies in each forest plot, which means that interpretations should be cautious.”

7. I wondered if effect sizes/statistical results should be included in second paragraph (starting ‘table 1…’) on page 16?

Response: The BMJ style guide recommends that results in tables are not duplicated in the text.

8. It was odd for table 1 to be the meta-regression findings. I would have expected at least a summary of results as the first table rather than an exploration of heterogeneity. Figure 2 should come before table 1 at least? Also, it is not immediately clear why models were not generated for some combinations in table 1?

Response: The summary of results is presented as Supplementary Table B (mentioned P8 and P11) as it was too large for the main body of the paper (11 pages). Even 1-2 lines on each study would result in a table being too large for the BMJ guidelines – but we are open to being further advised by the editors if they wish for us to present a summary table in the main body of the article. Figure 2 is presented on P14, ie before Table 1 is presented (end P15).

The models that were not generated have clarified this in the legend:

*non-linear, model not performed;

**no high risk populations;

***too few studies, I2 and τ2 not available

9. Table 3 – not immediately clear that this refers to quality scoring. Suggest tweak to title

Response: title changed to:

Table 3: Findings from agreement of quality rating analysis of coders using the Hawker tool

And P16: “Table 3 shows the quality ratings of the studies,..”

Reviewer: 2

General comments:

10. My main concern about this approach is whether there is any useful information to be gained by pooling data for this type of estimate. The logical follow up if administration of the 5 As is low would be to implement interventions to increase use of the 5 As. However, by pooling data, you are masking findings that would be useful for follow up, hiding the places where administration is low. The same is true for including years that aren’t recent. I'm concerned that data that are decades old don’t reflect current practice and so shouldn't be combined with recent data. Alternative approaches would be to restrict to recent years, examine countries or localities separately, and/or focusing on particular areas rather than all countries.

Response: The older papers (oldest is 1990) do not differ in any particular direction from more recent papers warranting exclusion. It was not our aim to only look at current practices only. The metaregression analysis was an opportunity to determine if date had an impact on heterogeneity, which did show up in the ‘Arrange Referral’ metaregression, and this has been commented on. The main problem for dividing up the analyses further by country, date range, localities, is that there would be very few papers then in each category.

11. Abstract

I found the non-standard abbreviations (such as HP and SCC) to be distracting rather than helpful. I would spell these out.

Response: now changed in abstract

12. Introduction

Page 5: there are more recent and more comprehensive resources documenting the effects of smoking on pregnancy outcomes. Suggest referencing the most recent Surgeon General’s reports and review articles. Some of the outcomes associated with prenatal smoking (like childhood cancers) are not established.

Response: We have cited the 2014 Surgeon General's report and removed the text 'childhood cancer'.

13. Line 40-41: less likely to abstain than whom?

Response: The comparators have been provided:

"than more advantaged women among whom smoking prevalence is lower" and "than non-pregnant women".

14. It isn't clear why the authors decided to look at NRT use in pregnant women in this analysis, given that many countries do not include a recommendation to use NRT in their clinical guidelines.

Response: Agreed, NRT is not recommended in all countries - we will address this point in the discussion as a limitation. However, all of the studies in the NRT 'yes' meta-analysis are US studies and there is a heterogeneity (from 11% to 47%) even though NRT is not recommended in the US for pregnancy. We will add this observation to the discussion, as follows:

P15 "All of the studies in the meta-analysis for 'Prescribing NRT – Yes' were from the USA (Figure Q supplementary file)."

P17 "However, all of the studies in the meta-analysis of NRT were from the USA, and considerable variation for prescribing NRT is seen within that one country."

P18 "We recognise that differing clinical guidelines may have impacted the provision of NRT in pregnancy in some countries. In particular NRT is not recommended for pregnancy in the USA."

15. Which countries recommend the 5 As vs. some other type of behavioral intervention?

Response: This study did not include a formal review of clinical guidelines from each country. However, the 5As are recommended by almost all of the countries. NZ recommends the ABC approach, and the UK recommends the AAA approach – all of these are based on the 5As so the first 2As are part of all clinical guidelines, and then some recommend cessation support and some recommend referral to cessation support.

We have added this as a limitation as follows:

P18 “Additionally, while most countries do use the 5As, there are variations, such as ABC (Ask, Brief Advice, Cessation) in NZ, and Ask, Advise, Action (AAA) in the UK. These have in common the first 2As, and then a variation to shorten the mnemonic or practice. This variation may be a limitation to this study.”

Methods

16. Suggest restricting to recent studies (see general comments).

Response: See previous response to #10. Furthermore we would like it to be noted that this was a systematic review for which we a priori decided the methodology and published it in PROSPERO. We would not like to spoil the rigour of the review by altering the methods post-hoc. Restricting the included studies to only the last 10 years would make the samples even smaller and further limit the meta-analyses.

17. Suggest restricting analysis of NRT to countries or localities where it is recommended for pregnant women.

Response: See response to #14. In addition, as per comment #16, we are following an a priori methodology. Numbers of studies describing NRT practices were too small to further divide the meta-analysis. For NRT yes/no only one paper was outside of the USA (ie in UK).

18. Page 8: How were studies of knowledge and attitudes and “other BCTs” incorporated into the analysis? The methods section here needs to be expanded.

Response: P8-9 added “and whether the study addressed the provision of BCTs, and if so a description of the BCTs (e.g., setting a quit date, increasing self-efficacy, monitoring carbon monoxide reading, validating abstinence).”

P9 added in relation to the narrative analysis: “including BCTs where reported.”

Results

19. Page 12: some of the outcomes come up unexpectedly. Is there any evidence base for measuring exhaled CO as a cessation strategy?

Response: On P8 we have expanded the list of BCTs that may have come up in the results (without pre-empting all of them). Thus have added:

“aiding social support, encouraging smoke-free environments,”

Measurement of exhaled CO is both a screening tool and a biofeedback technique that has been found to increase uptake of smoking cessation referrals in pregnancy and is now part of standard care in the UK. I have added a reference pertaining to this on P8.

20. Page 16: the description of the authors’ assessment of study quality (most had some aspects rated as good) does not give the reader an appreciation for the overall study quality of this body of literature. For example, what types of limitations were commonly found? How many studies were of high quality overall or in key areas? Inter-rater agreement wasn’t optimal. What did the authors do to reconcile this?

Response: 20 out of 53 (37.7%) studies that were rated had at least 5 ‘good’ categories out of the 9 available options. We have added this to the quality rating results on P16. Also added: “Common flaws were lack of clarity about aims, sampling processes not detailed, ethics processes not described, and no suggestions made for further research.”

We were unable to reconcile the rater’s agreement so have added on P18 “unresolved” to discrepancies between raters as a limitation.

Discussion

21. Page 20: There is not a strong evidence base for pharmacotherapy for pregnant women, with without psychosocial support. Therefore, it may not be bad or surprising that providers aren’t prescribing NRT.

Response: We acknowledge this interpretation.

Reviewer: 3

22. My main concern about this paper is the very high heterogeneity observed in the pooled prevalence rates (e.g., 99.1%). This very high heterogeneity represents a great variability among the rates reported by the examined studies, suggesting that the use of meta-analysis is questionable. Although the high heterogeneity is a very common issue in meta-analyses of prevalence rates, it is still not clear to me whether performing such meta-analyses is justifiable. The authors should consider the possibility of not pooling the rates of these very heterogeneous studies and focus on the narrative synthesis of the results. At the very least, the authors should try to explore what factors influence the variability of the prevalence rates and discuss the observed heterogeneity in details.

Response: The purpose of the meta-regressions was to explore any causes for heterogenicity. With the small number of studies, only a significant source of heterogeneity for 'Arrange Referral' only was found. With 54 studies in the review it would be unlikely that the heterogeneity could have been deciphered by a narrative synthesis alone, without making the paper unduly long. Despite the overall heterogenicity it is apparent that some elements of the 5As are more reliably performed and this can guide where to focus future interventions to improve SCC.

VERSION 2 – REVIEW

REVIEWER	Felix Naughton University of East Anglia, UK
REVIEW RETURNED	03-May-2019

GENERAL COMMENTS	The authors have addressed the issues raised or at least explained their take on these. The one thing I am not convinced by though is the rationale for undertaking fixed effects meta-analysis. I appreciate the reference to the Bornstein et al paper, which is helpful, but this paper does not recommend using a fixed effects model when studies are small in number, they merely indicate this as one undesirable option (among a number of undesirable options) as an alternative to random effects. They make efforts to say that if doing a fixed effects because the underlying distribution cannot be modelled, then the findings should not be considered generalisable and highlight that they will likely be read as such by readers even when saying they are not. Given the obvious heterogeneity in the studies I think this further indicates that fixed effects is not the way to go. But I am not a statistician and can only offer my opinion and will leave this decision to the editorial team.

VERSION 2 – AUTHOR RESPONSE

Response to reviewer:

The authors have addressed the issues raised or at least explained their take on these. The one thing I am not convinced by though is the rationale for undertaking fixed effects meta-analysis. I appreciate the reference to the Bornstein et al paper, which is helpful, but this paper does not recommend using a fixed effects model when studies are small in number, they merely indicate this as one undesirable option (among a number of undesirable options) as an alternative to random effects. They make efforts to say that if doing a fixed effects because the underlying distribution cannot be modelled, then the findings should not be considered generalisable and highlight that they will likely be read as such by readers even when saying they are not. Given the obvious heterogeneity in the studies I think this further indicates that fixed effects is not the way to go. But I am not a statistician and can only offer my opinion and will leave this decision to the editorial team.

Response: Given the problems in pooling effects when there are few studies, we believe a fixed effect approach is still worthwhile when there is overlap in confidence intervals from study specific effects, and believe these should be left in the results, but with a caution to the reader about the problems with pooling effects when there are a low number of studies. To be cautious we have removed discussion or presentation of any results in the text for estimates arising from n<5 studies.

Thus we have included this text in the discussion section P19-20:

“Where the number of studies was low (≤5), fixed effects modelling was used

because the between-studies variance (tau-squared), and therefore the mean of the underlying random distribution cannot be estimated with precision; heterogeneity is also not presented in these cases. We suggest these results are interpreted with caution, and consideration be given to the degree of overlap in the study specific confidence intervals.”