PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Comparative safety of anti-epileptic drugs for neurological development in children exposed during pregnancy and breastfeeding: a systematic review and network meta-analysis
AUTHORS	Veroniki, Areti Angeliki; Rios, Patricia; Cogo, Elise; Straus, Sharon; Finkelstein, Yaron; Kealey, M.; Reynen, Emily; Soobiah, Charlene; Thavorn, Kednapa; Hutton, Brian; Hemmelgarn, BR; Yazdi, Fatemeh; D'Souza, Jennifer; MacDonald, Heather; Tricco, Andrea

VERSION 1 - REVIEW

REVIEWER	Irene Petersen UCL, United Kingdom
REVIEW RETURNED	12-Apr-2017

GENERAL COMMENTS	This systematic review evaluate the risk of neurological outcome associated with antiepileptic drug treatment in pregnancy. A large Cochran review was recently published on the same topic (http://onlinelibrary.wiley.com/doi/10.1002/14651858.CD010236.pub2/full). Further details should be given to what this review add to the existing review. This review is based on observational studies, but unfortunately, many of the studies included in the review do not account for the underlying illnesses such as epilepsy and bipolar disorder. However, we know that there is a strong links between for example epilepsy and autism at individual level. It is therefore likely that women with epilepsy are of higher risk of giving birth to a child with autism irrespectively of treatments with AEDs. While there are increasing evidence suggesting that valproate treatment in pregnancy may be associated with adverse neurological childoutcomes. the issues of confounding (by indication and other factors) is of major concern in a systematic review solely based on observational studies. In particular, I am concerned about the findings for lamotrigine as this may, for many women, be the only treatment option left during pregnancy.

REVIEWER	Mervyn Eadie University of Queensland, Australia
REVIEW RETURNED	14-Apr-2017

GENERAL COMMENTS	This paper takes up a matter of considerable importance to women with antiepileptic drug treated seizure disorders, and to those responsible for their medical care. Its production has obviously

	involved a considerable amount of work in assembling and analysing data from previous studies in the area. The paper is written in rather terse prose and the main text is rather heavily involved in justifying the validity of the method of meta-analysis that has been employed, to some extent to the relative minimisation of the study's findings. It is possible that prospective readers who are not familiar with the papers methodology may have a little difficulty in following it unless they work their way through the rather numerous supplementary data. To obtain as much material as they could from various publications in the literature, the authors have had to omit a number of important studies in which findings have been presented in terms of continuous variables when the majority of the literature results have been expressed in terms of presence or absence of disturbance in particular aspects of neurodevelopment. The overall findings of the study are, not surprisingly, in general agreement with the findings of the original papers on which the meta-analysis is based. However, as the authors point out, their technique of meta-analysis has the additional advantage that it permits a potential ranking of culpability of individual antiepileptic drugs for disturbing neurodevelopment in utero and after birth. This is shown in Forest plots in the main body of the text, but the SUCRA analysis results appear only in the supplementary material. As well is this ranking, which the authors interpret reasonably critically in relation to the reliability of some of it where small numbers are involved, the paper serves the very useful purpose of bringing together all the relevant literature. By careful reading of all of the supplementary material it is possible to achieve a satisfactory understanding of what has been done in the study and what its possible limitations are. The authors have discussed a number of these but there are a few that they have not touched on, probably because no, or insufficient, appropriate data were available. Thus there is: • No consideration of a possible genetic contribution from the biological father, though information about this aspect would have involved issues of sensitivity • No clear indication of whether there were maternal seizures during pregnancy, though the issue of 'severity' of epilepsy is mentioned without clear indication of the meaning of 'severity' • No critical evaluation of the validity of the 'controls' in the various studies. While at first sight these appear to be entirely suitable, being the offspring of women with epilepsy that was not treated with antiepileptic drugs, this does raise the question of why these women were not treated and whether in some other way they may differ from treated populations. • Although the paper in the number of places refers to antiepileptic drug treatment during pregnancy or breastfeeding there does not seem to be any paper among those which provide the information for the present paper in which the drugs were used only during breastfeeding and not during pregnancy. This could be a matter of substantial clinical importance when the main culprit in relation to neurodevelopmental problems appears to be valproate. The practice in recent times has sometimes been adopted of ceasing that drug in anticipation of pregnancy to avoid foetal malformation, but resuming it in the second half of pregnancy. This policy may be unsafe in relation to neurodevelopment if later pregnancy and neonatal exposure to the drug is harmful from the neurodevelopmental

	standpoint. • When valproate is often considered the drug of choice for primary generalised epilepsies and appears to be the main culprit in relation to neurodevelopmental issues, the question arises as to whether there may be an association with this type of often inherited epilepsy and the neurodevelopmental issues I think it unlikely that the authors will be able to provide answer to the matters raised immediately above, but they might be touched on in the discussion section of the paper.

REVIEWER	Orestis Efthimiou University of Bern, Switzerland
REVIEW RETURNED	19-Apr-2017

GENERAL COMMENTS	My review mostly focuses on the statistical details of the analyses presented in this paper. The clinical context falls outside my area of expertise, so I cannot review the corresponding parts in the manuscript. Keeping that in mind, I think that overall the paper is very well written, its messages are clear and that it is methodologically sound. I have several comments, but most are rather minor or regard clarifications on the text. Title: 1. The title reads “Comparative neurological outcomes and safety of anti-epileptic rugs…”, but all analyses are about safety outcomes, as mentioned in the first sentence of the Abstract “Objectives: To compare the safety of Anti-epileptic drugs…” and also in the “Primary and secondary Outcome measures” of the abstract. So I do not really understand what are these neurological, non-safety outcomes that the title is referring at. Again, I am no expert in this medical field, if you think that my comment is irrelevant please disregard it. Abstract: 2. The results section writes “Results: The NMA on cognitive developmental delay 10 cohort studies, 748 children, 14 AEDs and control (no AED) suggested valproate …”. This reads weird, I think there is a verb missing in the first part, e.g. “The NMA on cognitive developmental delay INCLUDED 10 cohort studies … AND suggested”. 3. Later, same sentence “ … suggested valproate (arm sample size (N)=160)…” What is this “arm sample size”? Maybe you mean the total number of patients that received valproate? But later you also write “… and the combination carbamazepine, phenobarbital, and valproate (N=3…)” so it is probably not number of patients. Maybe number of different arms? But then again, 160 different arms in 748 patients doesn’t seem very probable, this is less than 5 patients per arm. Anyway, please clarify what this N is, or delete it altogether. 4. Later, same section, you write “were associated with a significantly greater risk of psychomotor delay compared with control.” You synthesized odds ratios, so throughout the manuscript you should talk about an increase in the odds, not an increase in the risk. 5. Conclusion section, it writes “Conclusions: Across all outcomes, valproate alone or combined with another AED is associated with the greatest risk, whereas…”. This is vague. Associated with the

	greatest risk of what? Also, compared with what (all other AEDs and combinations for example, or the control?). This ambiguity is also present in the article summary ("Across all neurological outcomes, valproate alone or combined with another AED is associated with the greatest risk.") and in the closing remark of this paper (page 23, line 437) Methods 6. Page 9, line 154 it writes "For each outcome with ≥10 studies and treatment comparisons with different total numbers of patients, the comparison-adjusted funnel plot was used to assess reporting bias, where the overall treatment effect for each comparison was estimated under the fixed-effect meta-analysis model." This is quite unclear. First, I do not understand the part "treatment comparisons with different total numbers of patients". What do you mean here by total number of patients? Also, if there were outcomes with ≥10 studies and treatment comparisons with EQUAL total numbers of patients you did not do a funnel plot? I don't get it… Second, a funnel plot does not (only) assess reporting bias. A funnel plot assesses the possible existence of small study effects (SSE), which encompasses publication bias, reporting bias, but also true differences in the relative effects between small and larger studies (eg. due to systematic differences in the studies populations). Consider rephrasing. Third, the only funnel plots you have presented are comparison-adjusted funnel plots, where you ordered the AEDs from newer to older. This comes a bit out of the blue. This way you may miss important SSE for AEDs vs. control. What I would do is present a regular (non-comparison adjusted…) funnelplot for all treatment comparisons that have enough studies. Then I would also put all AED vs. control (no AED) in a single funnel plot. This would allow the exploration of SSE in AEDs vs. control. Finally I would also present the comparison adjusted funnel plot, saying however that it only explores the hypothesis that newer AEDs are favoured over older ones (which is probably a much more underpowered analysis than the previous two). 7. Page 10, line 186, it writes "…assuming a common fixed coefficient across treatment comparisons". Probably here you mean a common fixed coefficient for all treatment comparisons for AEDs vs. the control, not actually ALL treatment comparisons. Please clarify. 8. Page 11, it writes "and higher methodological quality for the two items of the Newcastle-Ottawa Scale that had the highest percentage of low methodological quality (adequacy of follow-up of cohorts and comparability of cohorts items for cohort studies)." This sentence is rather hard to follow. Could you maybe rephrase? 9. Same page, lines 199-201, you discuss the use of DIC for comparing fit and parsimony of models. But, it is not clear from the context, which models do you aim to compare? In the results' section there is also no comparison between any models. 10. Page 12, line 216, it writes "SUCRA curve values are presented along with 95% CrIs to capture the uncertainty in the parameter values." There has been a discussion going on, on whether or not it makes sense to provide CrI for SUCRAs. For example this was discussed in the annual meeting of the society of research synthesis methods last year. The general consensus was that it doesn't make much sense to give CrIs for SUCRAs, mainly because SUCRAs incorporate uncertainty in their definition. You wouldn't give a CrI for a standard error or a p-value, so you shouldn't give a CrI for SUCRA

	as well. See also text in discussion lines 356 and 396 Results: 11. All the identified studies were cohort studies, so the methodological quality might vary a lot across the studies, and low quality studies might introduce a lot of bias. The section in the paper that present the results of this assessment is rather short and uninformative. Of course all information from the quality assessment is in Appendix I, but I think it would be much easier for the reader if the authors could provide some more detail in the main paper, section “methodological quality results”. For example, some points that might be of interest: how many studies controlled for confounding? What was the extent of missing outcome data in the original studies (e.g. how many patients were lost to follow up)? Did the studies adjust for this missing data? How many studies were deemed to be overall of high quality (low risk of bias), and how many were deemed to be of low quality? Etc. 12. In the statistical analysis results section it would be nice if you could provide the values for tau squared for each NMA, so that we can also see if the network meta-regressions explain some of the heterogeneity Discussion 13. Line 357, it writes “The probability that a top AED is actually among the worst one is likely high, Please rephrase: what is the worst one? Also the phrase “the probability is likely high” needs to be changed into “the probability is high”, or “it is likely that” 14. page 22, line 408, it writes “Recent research papers have explored methods to incorporate non-randomized with randomized evidence in a NMA and have highlighted the need to carefully explore the level of confidence in the non-randomized evidence (37, 38). However, the use of observational studies allows the assessment of the safety profile of AED treatments and offers the opportunity to evaluate effects in pregnancy”. I don’t think that these 2 papers you are citing provide any general argument against the use of observational evidence (as the word “however” in the quoted sentence seems to imply). What both these papers discuss is that when including non-randomized evidence you need to think about the confidence you want to place in the studies. For example, in this review 27 studies were identified, some of which were of high quality, some of low quality. Using the methodology of the 2 cited articles you could perform a NMA where you downweight studies depending on their quality, so that low quality studies would be allowed to have a smaller influence on the results. Please note that authors here should not feel obliged to do this kind of analysis due to this comment. I am not trying to take advantage of my reviewer status to promote this method – this would be inappropriate, given that I am the author of one of these papers. I am only trying to clarify how the methods in these 2 articles could be used here, and to point out that neither of these 2 articles argue against the use of observational studies, as your comment might imply.

VERSION 1 – AUTHOR RESPONSE

Reviewer 1

This systematic review evaluated the risk of neurological outcome associated with antiepileptic drug treatment in pregnancy. A large Cochrane review was recently published on the same topic (http://onlinelibrary.wiley.com/doi/10.1002/14651858.CD010236.pub2/full). Further details should be

given to what this review adds to the existing review.

Response: We discuss about the aforementioned review in the Discussion section, and to make clearer what our review adds to the previous one, we updated the following sentences as (lines 383-390):
“Also consistent with our results, a 2014 Cochrane review including 28 studies (10 of these studies were included in the meta-analyses; with a maximum number of five studies per meta-analysis) concluded that AED polytherapy led to poorer developmental outcomes and IQ compared to healthy controls, epileptic controls, and unspecified monotherapy.5 This Cochrane review also concluded that insufficient data exist for newer AEDs. However, unlike our review, it included and analysed fewer studies, and did not differentiate between specific polytherapy regimens, and thus did not compare these regimens versus each other or specific monotherapy AEDs.”

This review is based on observational studies, but unfortunately, many of the studies included in the review do not account for the underlying illnesses such as epilepsy and bipolar disorder. However, we know that there is a strong links between for example epilepsy and autism at individual level. It is therefore likely that women with epilepsy are of higher risk of giving birth to a child with autism irrespectively of treatments with AEDs.
While there are increasing evidence suggesting that valproate treatment in pregnancy may be associated with adverse neurological child outcomes. the issues of confounding (by indication and other factors) is of major concern in a systematic review solely based on observational studies. In particular, I am concerned about the findings for lamotrigine as this may, for many women, be the only treatment option left during pregnancy

Response: As noted in the results section, the assessment of transitivity across the treatment comparisons for treatment indication suggested no violation of the assumption. However, we agree that there may be confounding factors (known and unknown) which may impact our results. We updated the relevant sentence in the Discussion section as (lines 424-430):
“Fourth, although no intransitivity for most treatment effect modifiers assessed was evident, there was an imbalance in the methodological study quality appraisal across treatment comparisons and most outcomes, which may impact our results. Unknown factors or factors that could not be assessed due to dearth of data may pose the risk of residual confounding bias, and hence risk the validity of the transitivity assumption. However, the assessment of consistency suggested no disagreement between the different sources of evidence in the network.”

Reviewer 2
This paper takes up a matter of considerable importance to women with antiepileptic drug treated seizure disorders, and to those responsible for their medical care. Its production has obviously involved a considerable amount of work in assembling and analysing data from previous studies in the area. The paper is written in rather terse prose and the main text is rather heavily involved in justifying the validity of the method of meta-analysis that has been employed, to some extent to the relative minimisation of the study’s findings. It is possible that prospective readers who are not familiar with the papers methodology may have a little difficulty in following it unless they work their way through the rather numerous supplementary data.

Response: We agree with the reviewer that the reader of our paper needs to have some background on systematic reviews and network meta-analyses in order to fully comprehend the presentation of the methods and results. However, we follow relevant study recommendations in order to increase transparency of reporting and we provide all necessary information for the interested reader. As we mention in our paper, we followed the PRISMA extension for NMA reporting guidelines (see reference #13), the ISPOR for NMA guidelines to report on methods (see reference #12), as well as the Cochrane Handbook guidance on the methods conduction and reporting (see reference #44).

To obtain as much material as they could from various publications in the literature, the authors have had to omit a number of important studies in which findings have been presented in terms of continuous variables when the majority of the literature results have been expressed in terms of presence or absence of disturbance in particular aspects of neurodevelopment.

Response: Thank you for pointing this out. Our literature search identified 7 studies reporting continuous outcomes, and since the majority of the eligible studies reported the relevant information using dichotomous data, we decided to exclude the 7 studies from the systematic review. These studies are reported in Appendix C, as key excluded studies.

The overall findings of the study are, not surprisingly, in general agreement with the findings of the original papers on which the meta-analysis is based. However, as the authors point out, their technique of meta-analysis has the additional advantage that it permits a potential ranking of culpability of individual antiepileptic drugs for disturbing neurodevelopment in utero and after birth. This is shown in Forest plots in the main body of the text, but the SUCRA analysis results appear only in the supplementary material.

Response: Thank you for this comment. Indeed, one of the many advantages of applying a network meta-analysis is the presentation of the treatment hierarchy according to their efficacy or safety. Due to space limitations, we do not discuss the SUCRA results in the text, but we updated the information presented in the forest plots, including the SUCRA values.

As well is this ranking, which the authors interpret reasonably critically in relation to the reliability of some of it where small numbers are involved, the paper serves the very useful purpose of bringing together all the relevant literature. By careful reading of all of the supplementary material it is possible to achieve a satisfactory understanding of what has been done in the study and what its possible limitations are. The authors have discussed a number of these but there are a few that they have not touched on, probably because no, or insufficient, appropriate data were available. Thus there is:
• No consideration of a possible genetic contribution from the biological father, though information about this aspect would have involved issues of sensitivity
• No clear indication of whether there were maternal seizures during pregnancy, though the issue of ‘severity’ of epilepsy is mentioned without clear indication of the meaning of ‘severity’
• No critical evaluation of the validity of the ‘controls’ in the various studies. While at first sight these appear to be entirely suitable, being the offspring of women with epilepsy that was not treated with antiepileptic drugs, this does raise the question of why these women were not treated and whether in some other way they may differ from treated populations.
• Although the paper in the number of places refers to antiepileptic drug treatment during pregnancy or breastfeeding there does not seem to be any paper among those which provide the information for the present paper in which the drugs were used only during breastfeeding and not during pregnancy. This could be a matter of substantial clinical importance when the main culprit in relation to neurodevelopmental problems appears to be valproate. The practice in recent times has sometimes been adopted of ceasing that drug in anticipation of pregnancy to avoid foetal malformation, but resuming it in the second half of pregnancy. This policy may be unsafe in relation to neurodevelopment if later pregnancy and neonatal exposure to the drug is harmful from the neurodevelopmental standpoint.
• When valproate is often considered the drug of choice for primary generalised epilepsies and appears to be the main culprit in relation to neurodevelopmental issues, the question arises as to whether there may be an association with this type of often inherited epilepsy and the neurodevelopmental issues
I think it unlikely that the authors will be able to provide answer to the matters raised immediately above, but they might be touched on in the discussion section of the paper.

Response: Thank you for raising all these points. We have not collected information on the aforementioned items, because these were not (or were poorly) reported in the identified studies. In particular, we did not identify data on fathers, seizure frequency during pregnancy, exposure through breastfeeding only, types of epilepsy and family history of the mother. Also, most of the identified studies did not examine the validity of the controls, mainly because the control groups were always drawn from the same population as the intervention groups and were fairly comparable. In many studies, the women in the control groups were ones that had decided to stop taking AEDS during pregnancy. To capture all these points, we updated the Discussion section as shown below (lines 441-450):
"More evidence from long-term follow-up studies is required to further delineate neurodevelopmental risks in children. Future studies should assess the genetic contribution from the biological father, maternal seizures during pregnancy, exposure through breastfeeding only, types of epilepsy, and maternal family history. Registries should aim to include a suitable control group and collect information on potential confounders, such as alcohol and tobacco use, allowing researchers to identify the safest agents for different patient-level covariates, and enhance decision-making for healthcare providers and patients. A critical evaluation of the validity of the control group is also necessary, in order to examine potential differences between the treated and the not treated populations."

Reviewer 3
My review mostly focuses on the statistical details of the analyses presented in this paper. The clinical context falls outside my area of expertise, so I cannot review the corresponding parts in the manuscript. Keeping that in mind, I think that overall the paper is very well written, its messages are clear and that it is methodologically sound. I have several comments, but most are rather minor or regard clarifications on the text.
Title:
1. The title reads "Comparative neurological outcomes and safety of anti-epileptic rugs…", but all analyses are about safety outcomes, as mentioned in the first sentence of the Abstract "Objectives: To compare the safety of Anti-epileptic drugs…" and also in the "Primary and secondary Outcome measures" of the abstract. So I do not really understand what are these neurological, non-safety outcomes that the title is referring at. Again, I am no expert in this medical field, if you think that my comment is irrelevant please disregard it.

Response: Thank you for this comment. We updated the title as: "Comparative safety of anti-epileptic drugs for neurological development in children exposed during pregnancy and breastfeeding: a systematic review and network meta-analysis"

Abstract:
2. The results section writes "Results: The NMA on cognitive developmental delay 10 cohort studies, 748 children, 14 AEDs and control (no AED) suggested valproate …". This reads weird, I think there is a verb missing in the first part, e.g. "The NMA on cognitive developmental delay INCLUDED 10 cohort studies … AND suggested".

Response: The abstract has been corrected to read the following (lines 73-74):
"The NMA on cognitive developmental delay (11 cohort studies, 933 children, 17 AEDs and control) suggested among all AEDs only valproate…"

3. Later, same sentence " … suggested valproate (arm sample size (N)=160)…" What is this "arm sample size"? Maybe you mean the total number of patients that received valproate? But later you also write "… and the combination carbamazepine, phenobarbital, and valproate (N=3…)" so it is probably not number of patients. Maybe number of different arms? But then again, 160 different arms

in 748 patients doesn't seem very probable, this is less than 5 patients per arm. Anyway, please clarify what this N is, or delete it altogether.

Response: We understand the potential confusion, however your first assumption is correct in that "N" in both cases represent the number of individuals in each treatment arm. We deleted this detail in order to avoid confusion.

4. Later, same section, you write "were associated with a significantly greater risk of psychomotor delay compared with control." You synthesized odds ratios, so throughout the manuscript you should talk about an increase in the odds, not an increase in the risk.

Response: Thank you for this remark. We have updated the manuscript accordingly.

5. Conclusion section, it writes "Conclusions: Across all outcomes, valproate alone or combined with another AED is associated with the greatest risk, whereas…". This is vague. Associated with the greatest risk of what? Also, compared with what (all other AEDs and combinations for example, or the control?). This ambiguity is also present in the article summary ("Across all neurological outcomes, valproate alone or combined with another AED is associated with the greatest risk.") and in the closing remark of this paper (page 23, line 437)

Response: The conclusion in the abstract has been revised to the following (lines 85-86): "Valproate alone or combined with another AED is associated with the greatest odds of adverse neurodevelopmental outcomes compared with control."
We also updated the relevant bullet point as (lines 100-102:
"Across all neurological outcomes and treatments compared with control, valproate alone or combined with another AED is associated with the greatest odds of adverse development."
Similarly, in the Conclusions we updated the sentence to (lines 453-455): "Across all outcomes and treatments compared with control, valproate alone or combined with another AED was associated with the greatest odds, whereas oxcarbazepine and lamotrigine were associated with increased occurrence of autism."

Methods
6. Page 9, line 154 it writes "For each outcome with ≥10 studies and treatment comparisons with different total numbers of patients, the comparison-adjusted funnel plot was used to assess reporting bias, where the overall treatment effect for each comparison was estimated under the fixed-effect meta-analysis model." This is quite unclear. First, I do not understand the part "treatment comparisons with different total numbers of patients". What do you mean here by total number of patients? Also, if there were outcomes with ≥10 studies and treatment comparisons with EQUAL total numbers of patients you did not do a funnel plot? I don't get it… Second, a funnel plot does not (only) assess reporting bias. A funnel plot assesses the possible existence of small study effects (SSE), which encompasses publication bias, reporting bias, but also true differences in the relative effects between small and larger studies (eg. due to systematic differences in the studies populations). Consider rephrasing. Third, the only funnel plots you have presented are comparison-adjusted funnel plots, where you ordered the AEDs from newer to older. This comes a bit out of the blue. This way you may miss important SSE for AEDs vs. control. What I would do is present a regular (non-comparison adjusted…) funnelplot for all treatment comparisons that have enough studies. Then I would also put all AED vs. control (no AED) in a single funnel plot. This would allow the exploration of SSE in AEDs vs. control. Finally I would also present the comparison adjusted funnel plot, saying however that it only explores the hypothesis that newer AEDs are favoured over older ones (which is probably a much more underpowered analysis than the previous two).

Response: Thank you for this very thoughtful comment. We deleted the part "treatment comparisons

with different total numbers of patients" because we did not encounter such a case in our data. However, as suggested by the Cochrane Handbook, testing the funnel plot asymmetry when all studies are of similar sizes should be avoided. Similarly, study-specific treatment comparisons of similar sizes included in the comparison-adjusted funnel plot will have similar standard errors and the funnel plot will potentially not be very informative.
We agree, and re-phrased the term "publication bias" to "small-study effects".
We were unable to produce separate funnel plots for each treatment comparison and outcome, due to the small number of studies included in each NMA, and particularly in each treatment comparison (across all treatment comparisons and outcomes the maximum number of studies is 7; this is encountered in the Psychomotor Developmental Delay outcome).
Thank you for the suggestion on presenting regular funnel plots including all studies that contribute to a NMA, irrespective of the treatments they compare. Although this is a nice suggestion, we believe that there are two main disadvantages associated with this and we would prefer avoiding this approach. First, each treatment comparison comparing 2 specific treatments has its own overall mean effect, and testing asymmetry around different overall means would be challenging. The comparison-adjusted funnel plot has the advantage of centering all these means to zero. Second, data from multi-arm studies are correlated and a funnel plot assumes independency across all point estimates included in the plot. In the NMAs conducted in this manuscript, we have multiple multi-arm studies (25 of the total 29 studies) with number of arms ranging from 3 to 8 per trial (see Appendix J). Hence, our concern is that asymmetry can potentially be masked due to these correlations. This is also an issue in the comparison-adjusted funnel plot. To avoid such problems, we plotted only the study-specific basic parameters.
We decided to order the treatments chronologically so that the comparison adjusted funnel plot can be interpreted, as suggested by Chaimani et al (reference #16). We agree that this plot also assesses the hypothesis that newer AEDs are favoured over older ones, and we clarified this in the methods section (lines 167-170):
"All eligible medications were ordered from oldest to newest using their international market approval dates. Hence, the comparison-adjusted funnel plot additionally assesses the hypothesis that newer AEDs are favoured over older ones."

7. Page 10, line 186, it writes "…assuming a common fixed coefficient across treatment comparisons". Probably here you mean a common fixed coefficient for all treatment comparisons for AEDs vs. the control, not actually ALL treatment comparisons. Please clarify.

Response: We updated the relevant sentence accordingly.

8. Page 11, it writes "and higher methodological quality for the two items of the Newcastle-Ottawa Scale that had the highest percentage of low methodological quality (adequacy of follow-up of cohorts and comparability of cohorts items for cohort studies)." This sentence is rather hard to follow. Could you maybe rephrase?

Response: The sentence has been revised to read as follows (lines 205-208): "The sensitivity analysis for methodological quality was restricted to studies with low risk of bias for the two items on the NOS where the greatest proportion of studies received a low-quality score: adequacy of follow-up of cohorts and comparability of cohorts."

9. Same page, lines 199-201, you discuss the use of DIC for comparing fit and parsimony of models. But, it is not clear from the context, which models do you aim to compare? In the results' section there is also no comparison between any models.

Response: As noted in our methods section, we conducted network meta-regression analyses for maternal age and baseline risk when at least 10 studies were available. We aimed to compare the

DIC between the NMA model and the meta-regression models (given that the dataset analysed is the same across different models). However, we were able to apply only a single network meta-regression (cognitive developmental delay and for baseline risk), and hence we only report the DIC value for this outcome in the results section. All residual deviance and DIC values are reported in the appendices due to space limitations.
To clarify the models we aim to compare, we updated the relevant sentence as (lines 213-215):
"A difference of 3 units in the deviance information criterion between a NMA and a network meta-regression model was considered important and the lowest value of the deviance information criterion corresponded to the model with the best fit"

10. Page 12, line 216, it writes "SUCRA curve values are presented along with 95% CrIs to capture the uncertainty in the parameter values." There has been a discussion going on, on whether or not it makes sense to provide CrI for SUCRAs. For example this was discussed in the annual meeting of the society of research synthesis methods last year. The general consensus was that it doesn't make much sense to give CrIs for SUCRAs, mainly because SUCRAs incorporate uncertainty in their definition. You wouldn't give a CrI for a standard error or a p-value, so you shouldn't give a CrI for SUCRA as well. See also text in discussion lines 356 and 396

Response: We agree that SUCRAs take into account the uncertainty of the estimated treatment effects and this is mainly the reason we prefer presenting the SUCRA values instead of other ranking statistics (e.g., the probability of being the best). However, our NMAs include only a few studies ranging from 5 to 11 (with sample size ranging from 23 to 2011) across all outcomes, and most treatment comparisons are informed by a single study. Given that evidence show that the probability of being the best may be biased toward the treatments with the smallest number of studies (see Kibret et al reference #45), and that SUCRAs have a substantial degree of imprecision (see Trinquart et al # reference #31), we prefer presenting the 95% CrIs for the SUCRA values. Indeed, the 95% CrIs of the SUCRA values are very wide, indicating the high uncertainty around this estimation.

Results:
11. All the identified studies were cohort studies, so the methodological quality might vary a lot across the studies, and low quality studies might introduce a lot of bias. The section in the paper that present the results of this assessment is rather short and uninformative. Of course all information from the quality assessment is in Appendix I, but I think it would be much easier for the reader if the authors could provide some more detail in the main paper, section "methodological quality results". For example, some points that might be of interest: how many studies controlled for confounding? What was the extent of missing outcome data in the original studies (e.g. how many patients were lost to follow up)? Did the studies adjust for this missing data? How many studies were deemed to be overall of high quality (low risk of bias), and how many were deemed to be of low quality? Etc.

Response: Thank you for your comment about the methodological quality of the included studies. As noted in our paper, we have assessed all observational studies using the Newcastle Ottawa Scale and the results are reported in Appendix F. In the results section, we attempted to present the overall results of this assessment. To capture the points you raised, we updated the relevant sentences as shown below:
(lines 251-258): "Overall the studies were of good methodological quality and were rated as high quality across most items: 28 studies (97%) selected the non-exposed cohort from the same community as the exposed cohort, 26 (90%) included a representative or somewhat representative sample, 27 (93%) assessed outcomes independently, with blinding, or via a record linkage (e.g., identified through database records), and 23 (79%) ascertained exposure via secured records (e.g., database records) or structured interviews. The comparability of cohorts and adequacy of follow-up were the lowest scoring items across the studies with only 12 (41%) and 10 (34%) studies rated as high quality on these items."

12. In the statistical analysis results section it would be nice if you could provide the values for tau squared for each NMA, so that we can also see if the network meta-regressions explain some of the heterogeneity

Response: We have updated our results section accordingly.

Discussion
13. Line 357, it writes “The probability that a top AED is actually among the worst one is likely high, Please rephrase: what is the worst one? Also the phrase “the probability is likely high” needs to be changed into “the probability is high”, or “it is likely that”

Response: The aforementioned sentence has been deleted.

14. page 22, line 408, it writes “Recent research papers have explored methods to incorporate non-randomized with randomized evidence in a NMA and have highlighted the need to carefully explore the level of confidence in the non-randomized evidence (37, 38). However, the use of observational studies allows the assessment of the safety profile of AED treatments and offers the opportunity to evaluate effects in pregnancy”. I don’t think that these 2 papers you are citing provide any general argument against the use of observational evidence (as the word “however” in the quoted sentence seems to imply). What both these papers discuss is that when including non-randomized evidence you need to think about the confidence you want to place in the studies. For example, in this review 27 studies were identified, some of which were of high quality, some of low quality. Using the methodology of the 2 cited articles you could perform a NMA where you downweight studies depending on their quality, so that low quality studies would be allowed to have a smaller influence on the results. Please note that authors here should not feel obliged to do this kind of analysis due to this comment. I am not trying to take advantage of my reviewer status to promote this method – this would be inappropriate, given that I am the author of one of these papers. I am only trying to clarify how the methods in these 2 articles could be used here, and to point out that neither of these 2 articles argue against the use of observational studies, as your comment might imply.

Response: Thank you for this remark. Our intention was to highlight instances in which there is a need to use observational studies; one of which is the current clinical topic addressed in this paper. Therefore, we have deleted the word “However” to avoid confusion.
As a note, we agree with the reviewer that a NMA weighting studies depending on their quality would be very helpful to the reader. However, our NMAs include only a few number of studies, and in this case, such an analysis would have very low power. We will definitely consider this approach in a future study and update of the current systematic review if more studies are available.

VERSION 2 – REVIEW

REVIEWER	Mervyn Eadie Faculty of Health Sciences, University of Queensland, Brisbane, Australaia
REVIEW RETURNED	27-May-2017

GENERAL COMMENTS	I think the revisions made to this paper make it significantly clearer and stronger. There remain two quite minor points, viz. (i) In the Abstract there are a couple of instances where number of papers studied and number of subjects studied are not stated,

	whereas this information is otherwise usually provided. (ii) I have trouble with the specification 'in utero and/or breastfeeding' drug exposure in that there appear to be no data for breastfeeding without in utero exposure beforehand. It seems more a matter of ín utero exposure with or without subsequent exposure during breastfeeding.

REVIEWER	Orestis Efthimiou Institute of Social and Preventive Medicine, University of Bern, Switzerland
REVIEW RETURNED	29-May-2017

GENERAL COMMENTS	The authors addressed my comments satisfactorily. I only have one additional (very minor) comment to make. In discussion, it writes: "For example, the included studies often failed to report important confounding variables, such as family history of autism, ADHD, and maternal IQ, severity of epilepsy making it impossible for us to control these variables through subgroup analysis and meta regression." I disagree with this phrase, one cannot account for confounding by doing a subgroup analysis or a meta-regression. Controlling for confounding (i.e. estimation of causal rather than observational effects) can only be done at the study-level, when individual patient level are available. And anyway subgroup analysis and meta-regression constitute observational evidence on themselves (even if performed on RCTs), and they might lead to ecological biases. Please consider rephrasing. As it reads it might lead some readers to think that meta-regressing observational studies can remove confounding.

VERSION 2 – AUTHOR RESPONSE

Reviewer comments:

Reviewer 2
I think the revisions made to this paper make it significantly clearer and stronger.
There remain two quite minor points, viz.
(i) In the Abstract there are a couple of instances where number of papers studied and number of subjects studied are not stated, whereas this information is otherwise usually provided.
Response: Unfortunately, due to the limited space available for the abstract and the quantity of data, we are unable to provide detailed information for each treatment comparison. We have instead elected to report the number of included studies and number of infants/children included in our systematic review, as well as the number of studies, infants/children, and treatments analyzed for each outcome overall. However, if the editor feels that this information is necessary and the word count can be increased, we would be happy to include this information.

(ii) I have trouble with the specification 'in utero and/or breastfeeding' drug exposure in that there appear to be no data for breastfeeding without in utero exposure beforehand. It seems more a matter

of in utero exposure with or without subsequent exposure during breastfeeding.
Response: Thank you for raising an important point regarding clarity of reporting. The phrase you refer to appears in the Methods section of our manuscript where we describe the eligibility criteria for our review which, as stated in the review protocol (Additional File 1), consider for inclusion studies that report drug exposure only in utero, only during breastfeeding, or both. In the interest of brevity, we elected to describe these criteria using 'and/or' in our final manuscript. In contrast, throughout the Results section we have been careful to specify that in all of the included studies that report drug exposure through breastfeeding, this occurred subsequent to exposure in utero. Since only 5 of the studies included in our review reported exposure in utero and through breastfeeding and the remaining 24 studies only reported in utero exposure, we could not conduct any analyses based on this variable. To ensure that our reporting is as transparent as possible we have updated the table of Patient Characteristics (Appendix E) as well as the Summary Characteristics table (Table 1) to clearly indicate which of our included studies reported exposure in utero and through breastfeeding.

Reviewer 3
The authors addressed my comments satisfactorily. I only have one additional (very minor) comment to make. In discussion, it writes:
"For example, the included studies often failed to report important confounding variables, such as family history of autism, ADHD, and maternal IQ, severity of epilepsy making it impossible for us to control these variables through subgroup analysis and meta regression."
I disagree with this phrase, one cannot account for confounding by doing a subgroup analysis or a meta-regression. Controlling for confounding (i.e. estimation of causal rather than observational effects) can only be done at the study-level, when individual patient level are available. And anyway subgroup analysis and meta-regression constitute observational evidence on themselves (even if performed on RCTs), and they might lead to ecological biases.
Please consider rephrasing. As it reads it might lead some readers to think that meta-regressing observational studies can remove confounding.
Response:
Thank you for this comment! We agree with this point, and have updated the sentence as shown below (lines 415-418):
"For example, the included studies often failed to report important treatment effect modifiers,46 such as family history of autism, ADHD, and maternal IQ, severity of epilepsy making it impossible for us to explore their impact through subgroup analysis and meta-regression"