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)	The Impact of Including or Excluding Both-Armed Zero-Event Studies on Using Standard Meta-analysis Methods for Rare Event Outcome: A Simulation Study
AUTHORS	Cheng, Ji; Pullenayegum, Eleanor; Marshall, John; Iorio, Alfonso; Thabane, Lehana

VERSION 1 - REVIEW

REVIEWER	Fiona Warren University of Exeter Medical School (UEMS), UK
REVIEW RETURNED	01-Feb-2016

GENERAL COMMENTS	Review of: The Impact of Including or Excluding Both-Armed Zero-Event Studies on Using Standard Meta-analysis Methods for Rare Event Outcome: A Simulation Study This is an interesting paper, but is unfortunately difficult to read in places, and it is not always clear what the authors intended to express. The paper requires a technical edit for typos and grammar, and some rewriting to improve the clarity of English expression. I also have some technical and methodological concerns, set out below. I would like to see this paper published but it requires considerable revision both technically and editorially before it will meet publication standards. Abstract: 1. If using the abbreviation MA (which I am not keen on and prefer meta-analysis to be spelt out in full), it should be preceded by `an' not `a'. MA is pronounced Em Ay, so needs `an'. Would suggest spelling meta-analysis out in full throughout rather than abbreviating. 2. I'm not sure about the BAZE acronym. It may be simpler to refer to double-0 and single-0 studies. BAZE has been used in the context of Bayesian zero failure so to use the same acronym may be confusing. Strengths and limitations. 3. `reflect' not `reflex' in Point 2.

	Background 4. Suggest spelling SR out in full. See comment 1. 5. Para 2 first sentence. This concept is not clearly expressed. Suggest starting the para with something like: This research focuses on RCTs with a binary outcome (the participant does or does not experience the defined event). The total number of observed events in such an RCT is influenced by the event rate, sample size, and study period. 6. Para 2, final 2 sentences are grammatically incorrect. The final sentence implies that it is the zero-event study that has no events in either arm. Suggest something like: An extreme case of the zero-event study is the both-armed zero-event (double-0) study, which is defined as a study in which no event is observed in either the treatment or control arm, and is also known as a double-zero event or zero-total-event study. 7. Para 3 final sentence: state the n/N for studies that used 0.5 as the correction factor, as well as the percentage. 8. Para 5: suggest `….little empirical evidence…' rather than `…few empirical data…' 9. Para 5 final sentence: suggest saying `affect' rather than `moderate'. 10. Para 5: last three sentences could be restructured to more strongly state their hypotheses. For example: We hypothesize that the inclusion of double-0 studies may affect the pooled estimates of treatment effects in different ways, depending on the presence or absence of a true treatment effect. In the absence of a true treatment effect, i.e. the event rates are similar in both arms, the inclusion of double-0 studies may narrow the confidence interval of the pooled estimate of treatment effect. When a true treatment effect exists, inclusion of double-0 studies could affect the magnitude of the pooled estimate, leading to underestimation of the treatment effect. 11. I am not sure if I would consider the above as one hypothesis or two, ie one for the situation where there is not trued treatment effect, and one for where is a true treatment effect. It may be theoretically possible for one hypothesis to be correct but not the other, so it may be clearer to refer to two hypotheses rather than one. Also, it is not clear whether the authors are only investigating the situation where there is a true treatment effect, or both situations i.e. where is not a treatment effect as well as where there is. This needs to be clarified in the statement of what hypotheses are to be tested. (Given that the OR of 1 is included in the simulation parameters, the authors are investigating the situation with no treatment effect but this could be made clearer when stating the hypotheses.) 12. Para 6 `statistically deduce'. 13. The correct spelling is `Stata' not `STATA'. Method

14. Para 1, final sentence. This may imply that the only studies included in the meta-analysis were those with zero events in one arm. Maybe say: ….when studies with zero events in one, but not both, arms were included.

Simulation scenarios

15. Please state what the fixed, varied and derived parameters were.
16. Please define `low' and `high' impact parameters and explain how each parameter was classified as `low' or `high'.
17. Please state all formulae used to calculate the derived parameters (is this just the treatment arm event probabilities?).
18. The ORs where there is a treatment effect are expressed as <1 ie the treatment arm has fewer events than the control arm. It would be helpful to clarify whether the outcome is positive (eg patient cured) or negative (eg death). It is not clear whether the treatment arm is doing better than the control arm if OR<1. I assume that the treatment arm is expected to do better, therefore an OR<1 implies fewer negative events (ie outcome is negative). If the outcome is positive then the treatment arm is doing worse compared with controls. With an event rate of 0.001 and 5 studies of 100 patients each ie 500 patients per meta-analysis with 250 in the control groups, that leads to an expected number of events of 0.25 in the control arm and fewer in the treatment groups. Therefore, I would expect a high number of double-0 and single-0 studies. I would suggest that 0.001 is too low as a control group event rate and also that 5 is too low as the mean number of studies in a meta-analysis for the purposes of the simulations even if this is the median across all meta-analyses. The Sweeting et al study (2004) used 10 studies per meta-analysis and a control group event probability of 0.01 which may be more appropriate for simulation of sparse data. It may be helpful to use these parameters for the common scenario.
19. Rare events such as these are usually adverse events, and would usually be expected to be more frequent in the treatment arm, and less frequent in the control arm. It may have been helpful to express the ORs for treatment effect as >1, with the control group frequencies adjusted accordingly. This may make explanation of the results more intuitive, especially for the general readership of BMJ Open.
20. Para 2, final sentence: does this mean that in each simulation, all simulated meta-analyses included only simulated studies with the same number of patients? This is a possible methodological issue in that meta-analyses often combine smaller and larger studies. The effect of inclusion of double-0 studies may reflect the size of the double-0 studies, and inevitably, assuming the same event rate across all studies, it is the smaller studies that are more

likely to be double-0. This is a major methodological concern with the study (the authors have referred to this in the Discussion), and I believe it would improve the paper if the authors performed some further simulations to address the situation where there is variation in study size within a meta-analysis.

21. Para 3: Maybe say: the between study variation was applied at the level of the OR… .
22. Subheading: Methods for including both armed zero events studies… (not Methods to…). Add the `studies'.
23. The authors need to clarify their approach to continuity corrections for single-arm zero studies. I assume they used the 0.5 continuity correction for single-0 studies. The IV method requires a continuity correction for studies with 0 events in one arm only. The MH method only requires a continuity correction for single-arm zero studies if there are zero events in all trials in the meta-analysis for the same arm, but I think the metan command in Stata adds the continuity correction automatically if there are 0 events in one arm of a study. The Peto method does not require a continuity correction for single-zero studies.
24. Evaluating simulation performance: point 1 – state that it is the estimated/true value of the treatment effect rather than leaving it implicit.
25. Evaluating simulation performance: it is not clear what is meant by the percentage of number of studies included in the pooling process when excluding double-0 studies. Does this simply mean the average percentage of studies (where N=5) for each simulated meta-analysis that had zero events in both arms and are therefore excluded when double-0 studies are excluded? Or is it the proportion of studies out of 2500*5 that were double-0?
    This needs to be clarified (and in Table 2). The phrase "the inclusiveness of the approach of excluding double-0 studies in MA" is very unclear and needs to be clarified. It may also be worth reporting the percentage of studies with 0 events in only one arm, as this will impact on the method used to incorporate single-arm 0 event studies (when double-0 studies are excluded the bias will be driven by inclusion of single-0 studies). See also comment 36.

Table 1:

26. Column 3 heading: should be Rationale. Column 2: Many parameters have more than one assigned value so change column head to Assigned value(s).
27. Can I just check that row 4 refers to the total number of patients across the whole study not just per arm? This could be clarified, e.g. by saying Total number of patients… .

Table 2

28. Define symbols ie beta, delta, etc. Is beta the OR or log(OR)? Is delta the bias itself or the difference between the estimated OR and the true OR?

Statistical software and program

29. Please provide the R program for the simulations as an appendix.

Results

30. First paragraph should be in Methods section.
31. Please explain how the 57 simulated scenarios were derived. There are 13 parameters that vary across the simulations (4 ORs, 3 control arm event probabilities, 3 n patients per study, 3 between study SDs). There are 5 meta-analysis methods, and the scenarios are run with both inclusion and exclusion of double-0 studies. I am not clear how the 57 scenarios were derived.

Including double-0 studies

32. I am wondering if the results would be presented more clearly using a simple bar chart rather than the star charts (Figures 1 and 2).
33. Please present the results for RMSE, width of confidence intervals (CIs have a width not a length), and coverage in tables as appendices.
34. I would suggest presenting the bias on the log OR scale (see Sweeting et al 2004), for the advantage of having a symmetrical scale.
35. As a general comment on the structure of the results: the purpose of the paper is to provide guidance on whether to include or exclude double-0 studies in a meta-analysis. Therefore, what we need is a comparison of including and excluding double-0 studies across the four performance measures, as opposed to the current presentation, which is a narrative of the effect of inclusion of double-0 studies, followed by a narrative of the effects of excluding double-0 studies.

Excluding double-0 studies

36. Final sentence: does this mean that the proportion of double-0 studies increased (that were therefore excluded increased as OR decreased? How were the percentages derived? Are these the percentage of double-0 studies across all 2500 simulations of 5 studies each ie N=12,500? Given that each meta-analysis has 5 studies, I wonder if it might be relevant to provide a table indicating the mean number of double-0 studies per meta-analysis across the 2500 simulations for each scenario, and possibly the frequencies for the number of double-0 studies per meta-analysis, so we can get an idea of how many studies were being included in each meta-analysis. Were there any simulated meta-analyses where all studies were double-0? Also, the frequencies of single-0 studies may also be helpful.
37. The estimated ORs are consistently greater than the OR

used in the simulation (ie tending towards an OR=1), except when the OR in the simulation was 1. This would be expected when double-0 studies are included; when the double-0 studies are excluded, I assume the consistent bias towards the null is due to the inclusion of single-0 studies. It may be worth pointing this out in the discussion.

Peto method excluding double-0 studies

38. The Peto method produced the lowest bias across all scenarios where double-0 studies were excluded (except OR=1) but was only <0.8 for OR=0.8. This paragraph needs to be rewritten for clarity.

Summary

39. Suggest a Summary subheading to separate for previous paragraph.
40. To clarify, when OR =1 the other parameters remained unchanged in the simulation, so the statement `…regardless of the changes of control arm probability, number of patients and between-study variance.' Is not accurate as these parameter were not changed while holding OR constant.
41. It may also be helpful to clarify the direction of the bias, as well as stating that it was larger when double-0 studies were excluded, but in both cases (ie when including and excluding double-0 studies) the bias was in the same direction.

Tables 3a-3d

42. Should be `Taylor expansion'
43. Are some minus signs missing for OR=0.2?
44. Table 3a: as an example, if the true OR is 0.5 and the observed OR is 0.7 by the formula in Table 2, the bias would be positive not negative ie (0.7-0.5)/05 = 40% not-40% as stated. The ORs are being overestimated not underestimated (except when OR=1) so would expect bias to be positive. Have the authors deliberately changed the sign so that the bias is negative to express the concept that the bias is towards the null treatment effect? This is confusing as it does not reflect the formula in Table 2. It is unclear which is correct, the direction of the bias or the reported estimated ORs. It would also be easier to present bias on the log OR scale for symmetry of effect size.
45. Recurring typo in column heading: BA0E.
46. It would help the reader to state that a negative bias indicates a bias towards a positive effect of treatment (as outcome is negative), if this is the case.

Discussion

47. Para 1: bias was negative so inclusion increased the magnitude of the bias.

	48. Para 2: see point 17 above. Need to make clear that this study looked for a negative outcome therefore the conservative estimates favour the null hypothesis and would fail to pick up on a positive treatment effect. 49. Para 4, first sentence: maybe make more explicit that the bias for both excluding double-0 studies but including single-0 studies, and for including both types of study, was towards the null effect and therefore would underestimate a negative treatment effect. 50. Para 5: see comment 18 above with regard to the same number of patients in all studies. Also the number of studies per meta-analysis and control event rate may be too low. 51. Para 6: make clear that the Poisson regression method can incorporate double-0 studies (ie not just zero events in one arm only) without a continuity correction. Does the Spittal study include double-0 studies or just single-0 studies? Same for Bayesian methods. Conclusion 52. First sentence does not make sense. To say "…when treatment effects are unlikely preserve data integrality of the systematic review." has no clear meaning and I cannot advise on rephrasing as I have no idea what the authors intended to express. 53. Second sentence: not sure if `safer' is the right concept – maybe say `introduce less bias compared with alternative approaches'. This can link in with the sentence about the purpose of the review, ie less likely to underestimate a positive effect (where treatment is in fact beneficial) or underestimate a negative effect (where treatment is in fact harmful compared with control). Figures 54. Peto without BAZE is included twice in both figures.

REVIEWER	Oliver Kuss German Diabetes Center Leibniz Institute for Diabetes Research at Heinrich Heine University Düsseldorf, Institute for Biometry and Epidemiology, Düsseldorf, Germany
REVIEW RETURNED	05-Feb-2016

GENERAL COMMENTS	Cheng et al. report a simulation study to compare meta-analytic methods for binary outcomes, when studies without any event in both treatment arms (BAZE) are available. The paper is written well, concise and short, also the simulation study is well performed. I especially liked that the authors chose realistic and empirically verified values for the true parameters in the simulation. Unfortunately, the authors miss the most complete study in this field which is that of Kuss(2015). I argued, that enough meta-analytic

	methods that include BAZE studies are available and therefore (and also for other reasons) other methods using continuity corrections should no longer been used. I collected a number of arguments for that, the most convincing for me is the ethical one, originally raised by Keus: Patients who have been recruited to BAZE studies have a right to their data being also included in meta-analyses and this brings the duty for the researcher to include BAZE studies. I also included some of the methods from Cheng et al., and all of them were found inferior as compared to the finally recommended method, a beta-binomial regression model. As such, I am sorry to say that the paper of Cheng et al. delivers no relevant additional information and should be rejected. One could argue, especially in current days where replication is hotly debated, that the Cheng study could be seen as a welcome replication of my study. If yes, then Cheng et al. should definitely include the beta-binomial method in their simulation. Some minor points: - It is hard for me to accept with today's computing power that Cheng et al. only run 2.500 simulations, especially in their scenarios with only small meta-analyses. I feel 10.000 runs is a minimum here and a sensible compromise between computing time and random variation in simulation results. - It was not clear for me how the true meta-analyses were generated. It felt like they were generated from a standard inverse-variance random effects model, but the authors could be more clear here. - It is a good tradition in methodical papers in biostatistics, that the authors illustrate their problem with an example from real data. Why did Cheng et al. not use this option? Reference: Kuss O. Statistical methods for meta-analyses including information from studies without any events-add nothing to nothing and succeed nevertheless. Stat Med. 2015 Mar 30;34(7):1

VERSION 1 – AUTHOR RESPONSE

Responses to Fiona Warren's Comments

This is an interesting paper, but is unfortunately difficult to read in places, and it is not always clear what the authors intended to express. The paper requires a technical edit for typos and grammar, and some rewriting to improve the clarity of English expression. I also have some technical and methodological concerns, set out below. I would like to see this paper published but it requires considerable revision both technically and editorially before it will meet publication standards.
Abstract:
If using the abbreviation MA (which I am not keen on and prefer meta-analysis to be spelt out in full), it should be preceded by `an' not `a'. MA is pronounced Em Ay, so needs `an'. Would suggest spelling meta-analysis out in full throughout rather than abbreviating.
Response: We appreciate the reviewer for pointing out this detail. We have changed all "MA"s to meta-analysis throughout the manuscript.

2. I'm not sure about the BAZE acronym. It may be simpler to refer to double-0 and single-0 studies. BAZE has been used in the context of Bayesian zero failure so to use the same acronym may be

confusing.
Response: We thank the review for pointing this out. To avoid confusion, we have changed “BAZE” to “BA0E” to refer “both-armed zero-event” in the manuscript.

Strengths and limitations.
3 `reflect' not `reflex' in Point 2.
Response: We have corrected this typo.

Background
4. Suggest spelling SR out in full. See comment 1.
Response: We have replaced “SR” with “systematic review” throughout the manuscript

5. Para 2 first sentence. This concept is not clearly expressed. Suggest starting the para with something like: This research focuses on RCTs with a binary outcome (the participant does or does not experience the defined event). The total number of observed events in such an RCT is influenced by the event rate, sample size, and study period.
Response: We have restructured the first two sentences as suggested. (Page 6)

6. Para 2, final 2 sentences are grammatically incorrect. The final sentence implies that it is the zero-event study that has no events in either arm. Suggest something like: An extreme case of the zero-event study is the both-armed zero-event (double-0) study, which is defined as a study in which no event is observed in either the treatment or control arm, and is also known as a double-zero event or zero-total-event study.
Response: We have restructured this sentences as suggested. (Page 7)

7. Para 3 final sentence: state the n/N for studies that used 0.5 as the correction factor, as well as the percentage.
Response: The number of studies used 0.5 as the correction factor has been added. (Page 8)

8. Para 5: suggest `….little empirical evidence…' rather than `…few empirical data…'
Response: We have restructured the sentence accordingly. (page 8 -9)

9. Para 5 final sentence: suggest saying `affect' rather than `moderate'.
Response: We have replaced “Moderate” with “affect”. (Page 10)

10. Para 5: last three sentences could be restructured to more strongly state their hypotheses. For example: We hypothesize that the inclusion of double-0 studies may affect the pooled estimates of treatment effects in different ways, depending on the presence or absence of a true treatment effect. In the absence of a true treatment effect, i.e. the event rates are similar in both arms, the inclusion of double-0 studies may narrow the confidence interval of the pooled estimate of treatment effect. When a true treatment effect exists, inclusion of double-0 studies could affect the magnitude of the pooled estimate, leading to underestimation of the treatment effect.
Response: We thank the reviewer for the input. The sentences have been rephrased as suggested. (Page 10)

11. I am not sure if I would consider the above as one hypothesis or two, ie one for the situation where there is not trued treatment effect, and one for where is a true treatment effect. It may be theoretically possible for one hypothesis to be correct but not the other, so it may be clearer to refer to

two hypotheses rather than one. Also, it is not clear whether the authors are only investigating the situation where there is a true treatment effect, or both situations i.e. where is not a treatment effect as well as where there is. This needs to be clarified in the statement of what hypotheses are to be tested. (Given that the OR of 1 is included in the simulation parameters, the authors are investigating the situation with no treatment effect but this could be made clearer when stating the hypotheses.)
Response: We agree with the reviewer for her concern. We have revised the sentence to reflect two scenarios under the same topic: the presence and absence of a true treatment effect. (Page 10)

12. Para 6 `statistically deduce'.
Response: The typo has been corrected. (Page 11)

13. The correct spelling is `Stata' not `STATA'.
Response: STATA has been replaced by Stata

Method

14. Para 1, final sentence. This may imply that the only studies included in the meta-analysis were those with zero events in one arm. Maybe say: ….when studies with zero events in one, but not both, arms were included.
Response: We have restructured the sentence as suggested. (Page 11)

Simulation scenarios

15. Please state what the fixed, varied and derived parameters were.
Response: The definitions of fixed, varied and derived parameters have been added. (Page 11)

16. Please define `low' and `high' impact parameters and explain how each parameter was classified as `low' or `high'.
Response: We have described the "high" and "low" impacts in the paragraph. (Page 11)

17. Please state all formulae used to calculate the derived parameters (is this just the treatment arm event probabilities?).
Response: In our simulation, only the treatment arm probability was derived from the given values of the controlled arm probability and treatment effect. Therefore, only one formula was provided in the manuscript.

18. The ORs where there is a treatment effect are expressed as <1 ie the treatment arm has fewer events than the control arm. It would be helpful to clarify whether the outcome is positive (eg patient cured) or negative (eg death). It is not clear whether the treatment arm is doing better than the control arm if OR<1. I assume that the treatment arm is expected to do better, therefore an OR<1 implies fewer negative events (ie outcome is negative). If the outcome is positive then the treatment arm is doing worse compared with controls. With an event rate of 0.001 and 5 studies of 100 patients each ie 500 patients per meta-analysis with 250 in the control groups, that leads to an expected number of events of 0.25 in the control arm and fewer in the treatment groups. Therefore, I would expect a high number of double¬0 and single-0 studies. I would suggest that 0.001 is too low as a control group event rate and also that 5 is too low as the mean number of studies in a meta-analysis for the purposes of the simulations even if this is the median across all meta-analyses. The Sweeting et al study (2004) used 10 studies per meta-analysis and a control group event probability of 0.01 which may be more appropriate for simulation of sparse data. It may be helpful to use these parameters for the common scenario.
Response: We appreciate the reviewer's insight. With the combination of the extremely low event rates and small numbers of studies and patients in each study, the number of events in our simulation

indeed was very low. We chose the values of the simulation parameters to reflex the common characteristics of meta-analyses (including published and unpublished studies) for rare event outcomes. In Davey J and colleagues' review (BMC Res Methodol. 2011), among 14,886 meta-analysis for binary outcomes, the median (Q1, Q3) number of studies included in meta-analyses was 3 (2, 6); the median (Q1, Q3) number of patients in individual studies was 102 (50, 243). The event rates (0.1% to 1%) were chose according to the definitions of rare diseases: less 1 in 2000 in EU (EURODIS, 2005); rare adverse events: less than 1 in 1000 in US (Institution of Medicine, 2007). To sufficiently demonstrate the potential bias in the results of the small meta-analyses for rare event outcomes, in which both-armed zero-events are often presented, we feel that keeping the current settings of our simulations will serve the purpose of our study.

19. Rare events such as these are usually adverse events, and would usually be expected to be more frequent in the treatment arm, and less frequent in the control arm. It may have been helpful to express the ORs for treatment effect as >1, with the control group frequencies adjusted accordingly. This may make explanation of the results more intuitive, especially for the general readership of BMJ Open.
Response: We agree with the reviewer's suggestion. In the previous version of the manuscript, we discussed the potential impact of underestimating treatment effect when both-armed zero-event studies were included in meta-analyses, in which the outcomes could be beneficial or harmful events, but our simulation only showed on the direction of risk reduction. To make the simulation more intuitive to the readers, we have added three more simulation scenarios with the treatment effects set as odds ratios equal to 1.25, 2 and 5, which mimicked the scenarios of the increase of adverse effects in treatment arms. (Page 14 and Table 3a)

20. Para 2, final sentence: does this mean that in each simulation, all simulated meta-analyses included only simulated studies with the same number of patients? This is a possible methodological issue in that meta-analyses often combine smaller and larger studies. The effect of inclusion of double-0 studies may reflect the size of the double-0 studies, and inevitably, assuming the same event rate across all studies, it is the smaller studies that are more likely to be double-0. This is a major methodological concern with the study (the authors have referred to this in the Discussion), and I believe it would improve the paper if the authors performed some further simulations to address the situation where there is variation in study size within a meta-analysis.
Response: We agree with the reviewer that using the same number of patients for each individual study in a meta-analysis was a simplified approach and this setting barely reflect the real-word meta-analyses. However, because both-armed zero-events are likely to be observed in small studies, using varied numbers of patients in the simulation might cause the impact of including both-armed zero-events over shadowed by the large studies due to the weighting strategy. Therefore, to emphasize the potential bias may introduced by including both-armed zero-event studies in meta-analyses, we chose using the same number of patients for each individual studies in our simulation.

21. Para 3: Maybe say: the between study variation was applied at the level of the OR… .
Response: We thank the reviewer for the suggestions. The sentence has been revised accordingly. (Page 14)

22. Subheading: Methods for including both armed zero events studies… (not Methods to…). Add the `studies'.
Response: The subheading has been revised accordingly. (Page 15)

23. The authors need to clarify their approach to continuity corrections for single-arm zero studies. I assume they used the 0.5 continuity correction for single-0 studies. The IV method requires a continuity correction for studies with 0 events in one arm only. The MH method only requires a continuity correction for single-arm zero studies if there are zero events in all trials in the meta-

analysis for the same arm, but I think the metan command in Stata adds the continuity correction automatically if there are 0 events in one arm of a study. The Peto method does not require a continuity correction for single-zero studies.
Response: The clarifications have been added in this paragraph regarding to the continuity corrects. (Page 15)

24. Evaluating simulation performance: point 1 – state that it is the estimated/true value of the treatment effect rather than leaving it implicit.
Response: We have taken the reviewers suggestion and revised the sentence. (Page 16)

25. Evaluating simulation performance: it is not clear what is meant by the percentage of number of studies included in the pooling process when excluding double-0 studies. Does this simply mean the average percentage of studies (where N=5) for each simulated meta-analysis that had zero events in both arms and are therefore excluded when double-0 studies are excluded? Or is it the proportion of studies out of 2500*5 that were double-0? This needs to be clarified (and in Table 2). The phrase "the inclusiveness of the approach of excluding double-0 studies in MA" is very unclear and needs to be clarified. It may also be worth reporting the percentage of studies with 0 events in only one arm, as this will impact on the method used to incorporate single-arm 0 event studies (when double-0 studies are excluded the bias will be driven by inclusion of single-0 studies). See also comment 36.
Response: We thank the reviewer for pointing this out. We realized that the "inclusiveness" we defined previously as a measure of evaluating simulation performance added confusions. It can be simple expressed as the percentage of the number of studies with both-armed zero-events. Therefore, we have removed this term from Table 2 (Measures for evaluating simulation performance) and any related contents in method section. Instead, we simply reported the percentage of the number of both-armed zero-event studies in the result section. (Table 1, Page 19).

Table 1:
26. Column 3 heading: should be Rationale. Column 2: Many parameters have more than one assigned value so change column head to Assigned value(s).
Response: Reviewer's suggestion has been taken and the typo has been corrected. (Table 1)

27. Can I just check that row 4 refers to the total number of patients across the whole study not just per arm? This could be clarified, e.g. by saying Total number of patients… .
Response: Since the number of events (i.e. the chance of being both-armed zero-event study) in each individual study influenced by the number of patients in that study and other factors, we chose the report the number of patients in each individual study.

Table 2
28. Define symbols ie beta, delta, etc. Is beta the OR or log(OR)? Is delta the bias itself or the difference between the estimated OR and the true OR?
Response: The symbols used in Table 2 have been proper defined as footnotes under the table (Page 17).

29. Please provide the R program for the simulations as an appendix.
Response: R program has been provided as Appendix 1

Results
30. First paragraph should be in Methods section
Response: We have merged this paragraph to the method section. (Page 18)

31. Please explain how the 57 simulated scenarios were derived. There are 13 parameters that vary

across the simulations (4 ORs, 3 control arm event probabilities, 3 n patients per study, 3 between study SDs). There are 5 meta-analysis methods, and the scenarios are run with both inclusion and exclusion of double-0 studies. I am not clear how the 57 scenarios were derived.
Response: We thank the reviewer for pointing out this discrepancy. We initially created 57 simulation scenarios by setting more varied values, but only chose to reported 13 scenarios based on the focus of our investigation. We have corrected the reporting error. (Page 19)

Including double-0 studies
32. I am wondering if the results would be presented more clearly using a simple bar chart rather than the star charts (Figures 1 and 2).
Response: We agree with the reviewer that the star charts might not be the most intuitive graphical presentation for some readers. However, Figures 1 (RMSE) and 2 (width of 95% CI) were continuous numbers, which might not be suitable to use bar-chart since the bar chart is typically used to display proportions or percentages. To improve the quality of these two figures, we have enlarged the plots.

33. Please present the results for RMSE, width of confidence intervals (CIs have a width not a length), and coverage in tables as appendices
Response: We have added three tables for RMSE, width of CIs and Coverage as appendices. (Appendix 3a,3b, 3c)

34. I would suggest presenting the bias on the log OR scale (see Sweeting et al 2004), for the advantage of having a symmetrical scale.
Response: We presented the bias in OR scale according to the suggestions of most clinicians we worked with. However, we felt the reviewer's suggestion has the valid point. Therefore, we added the bias in log OR scale as Appendix 2.

35. As a general comment on the structure of the results: the purpose of the paper is to provide guidance on whether to include or exclude double-0 studies in a meta-analysis. Therefore, what we need is a comparison of including and excluding double-0 studies across the four performance measures, as opposed to the current presentation, which is a narrative of the effect of inclusion of double-0 studies, followed by a narrative of the effects of excluding double-0 studies.
Response: We realized that the structure used to present the results may not be the most natural way of expression. However, the objective of our study is eventually created some general strategies to inform the readers how to deal with both-armed zero-event studies in two aspects: including and excluding them. We feel that presenting the results of including and excluding both-armed zero-event studies in separate paragraphs is easier for us to make our argument in the later paragraphs.

Excluding double-0 studies
36. Final sentence: does this mean that the proportion of double-0 studies increased (that were therefore excluded increased as OR decreased? How were the percentages derived? Are these the percentage of double-0 studies across all 2500 simulations of 5 studies each ie N=12,500? Given that each meta-analysis has 5 studies, I wonder if it might be relevant to provide a table indicating the mean number of double-0 studies per meta-analysis across the 2500 simulations for each scenario, and possibly the frequencies for the number of double-0 studies per meta-analysis, so we can get an idea of how many studies were being included in each meta-analysis. Were there any simulated meta-analyses where all studies were double-0? Also, the frequencies of single-0 studies may also be helpful.
Response: The reviewer also pointed out this problem in comment #25. We realized that the inclusiveness simply meant the percentage of none both-armed zero-event studies. Using the term of inclusiveness only added confusion. Therefore, we have removed this term and simply reported the percentage of both-armed zero-event studies in the simulated data sets in the beginning of the result section.

37. The estimated ORs are consistently greater than the OR used in the simulation (ie tending towards an OR=1), except when the OR in the simulation was 1. This would be expected when double-0 studies are included; when the double-0 studies are excluded, I assume the consistent bias towards the null is due to the inclusion of single-0 studies. It may be worth pointing this out in the discussion.
Response: We thank the reviewer for the suggestion. We added some discussions regarding to this observation. In order to discuss this issue in depth, we feel more studies are needed. Since this is not main our study objective in this simulation, we will leave it a future work. (Page 27)

38. The Peto method produced the lowest bias across all scenarios where double-0 studies were excluded (except OR=1) but was only <0.8 for OR=0.8. This paragraph needs to be rewritten for clarity.
Response: We thank the reviewer for pointing this out. We have rewritten this paragraph with more clarity. (Page24)

Summary
39. Suggest a Summary subheading to separate for previous paragraph.
Response: The subheading has been added as suggested by the reviewer. (Page25)

40. To clarify, when OR =1 the other parameters remained unchanged in the simulation, so the statement `…regardless of the changes of control arm probability, number of patients and between-study variance.' Is not accurate as these parameter were not changed while holding OR constant.
Response: The sentence has been rewritten with more clarity. (Page 25)

41. It may also be helpful to clarify the direction of the bias, as well as stating that it was larger when double-0 studies were excluded, but in both cases (ie when including and excluding double-0 studies) the bias was in the same direction.
Response: The sentence has been rewritten to reflect the direction of the bias, which was underestimating the treatment effect. (Page 25)

Tables 3a-3d
42. Should be `Taylor expansion'
Response: We have redefined the formulas using a more direct approach for calculating the percentage bias in OR scale (Table 3a-3d) and absolute bias in log OR scale (Appendix 2). The Taylor expansion has been removed from the content.

43. Are some minus signs missing for OR=0.2?
Response: We thank reviewer to point out this error. The minus signs have been added to where they were missed.

44. Table 3a: as an example, if the true OR is 0.5 and the observed OR is 0.7 by the formula in Table 2, the bias would be positive not negative ie (0.7-0.5)/05 = 40% not-40% as stated. The ORs are being overestimated not underestimated (except when OR=1) so would expect bias to be positive. Have the authors deliberately changed the sign so that the bias is negative to express the concept that the bias is towards the null treatment effect? This is confusing as it does not reflect the formula in Table 2. It is unclear which is correct, the direction of the bias or the reported estimated ORs. It would also be easier to present bias on the log OR scale for symmetry of effect size.
Response: We thank the reviewer for pointing out this issue in details. We have redefined the formulas for calculating the percentage bias in OR scale (Table 3a-3d) and the absolute bias in log OR scale (Appendix 3). We also added the statement to make it clear to the readers that the

“negative (-)” sign associated with the bias measurement meant under estimating treatment effect i.e. pulling the estimates of the treatment effect towards the null hypothesis, which is OR =1 (log OR = 0).

45. Recurring typo in column heading: BA0E.
Response: We have changed the abbreviation from BAZE to BA0E to refer both-armed zero-event study in the entire manuscript to avoid confusions.

46. It would help the reader to state that a negative bias indicates a bias towards a positive effect of treatment (as outcome is negative), if this is the case.
Response: Statements have been added for Table 3a-d and Appendix 3 to make it clear to the readers that the negative signs associated with bias meant under estimating the true treatment effect.

Discussion
47. Para 1: bias was negative so inclusion increased the magnitude of the bias.
Response: We have stated the direction of the bias to make it clear that including both-armed zero-event studies added bias towards underestimating treatment effect. (Page 26)

48. Para 2: see point 17 above. Need to make clear that this study looked for a negative outcome therefore the conservative estimates favour the null hypothesis and would fail to pick up on a positive treatment effect
Response: We have made the statements to clarify the bias was in the direction of underestimating treatment effects (Pages 19, 24, 25, 26, Table 3a-3d and Appendix 3).

49. Para 4, first sentence: maybe make more explicit that the bias for both excluding double-0 studies but including single-0 studies, and for including both types of study, was towards the null effect and therefore would underestimate a negative treatment effect.
Response: We thank the reviewer for this suggestion. We have added more contexts to make the statement clearer to the readers. (Page 27-28)

50. Para 5: see comment 18 above with regard to the same number of patients in all studies. Also the number of studies per meta-analysis and control event rate may be too low.
Response: As explained in Comment #18. The event rates were chosen according to definitions of rare diseases or rare adverse events. In regarding to the number of patients used in this simulation, they seem low comparing to the meta-analyses published in the high impact journals, but there are a lot of small meta-analyses conducted as internal studies for policy makers. As reported in Davey J and colleagues’ review (BMC Res Methodol. 2011), among 14,886 meta-analysis for binary outcomes, the median (Q1, Q3) number of studies included in meta-analyses was 3 (2, 6); the median (Q1, Q3) number of patients in individual studies was 102 (50, 243). We felt that the choice of those values serves our study objective. Since both-armed zero-events are likely to be observed in underpowered studies or for secondary outcomes, we attempt to point out to the readers that including those both-armed zero-event studies in meta-analyses could lead to bias in underestimating treatment effects.

51. Para 6: make clear that the Poisson regression method can incorporate double-0 studies (ie not just zero events in one arm only) without a continuity correction. Does the Spittal study include double-0 studies or just single-0 studies? Same for Bayesian methods.
Response: We have added the clarification in the sentence. (Page 30)

52. First sentence does not make sense. To say “…when treatment effects are unlikely preserve data integrality of the systematic review.” has no clear meaning and I cannot advise on rephrasing as I have no idea what the authors intended to express.

Response: We have revised this paragraph to make explicit conclusions according the comments from the reviewers (including some internal reviewers). We created some strategies on dealing with both-armed zero-events studies for using standard meta-analysis methods for rare event outcomes. (Page 31-32)

53. Second sentence: not sure if `safer' is the right concept – maybe say `introduce less bias compared with alternative approaches'. This can link in with the sentence about the purpose of the review, ie less likely to underestimate a positive effect (where treatment is in fact beneficial) or underestimate a negative effect (where treatment is in fact harmful compared with control).
Response: We have revised the conclusion. (page 31-32).

Figures
54. Peto without BAZE is included twice in both figures.
Response: This typo has been corrected

Responses to Oliver kuss's Comments

Cheng et al. report a simulation study to compare meta-analytic methods for binary outcomes, when studies without any event in both treatment arms (BAZE) are available. The paper is written well, concise and short, also the simulation study is well performed. I especially liked that the authors chose realistic and empirically verified values for the true parameters in the simulation.
Response: We thank the reviewer for the encouraging comments.

Unfortunately, the authors miss the most complete study in this field which is that of Kuss(2015). I argued, that enough meta-analytic methods that include BAZE studies are available and therefore (and also for other reasons) other methods using continuity corrections should no longer been used. I collected a number of arguments for that, the most convincing for me is the ethical one, originally raised by Keus: Patients who have been recruited to BAZE studies have a right to their data being also included in meta-analyses and this brings the duty for the researcher to include BAZE studies. I also included some of the methods from Cheng et al., and all of them were found inferior as compared to the finally recommended method, a beta-binomial regression model.
As such, I am sorry to say that the paper of Cheng et al. delivers no relevant additional information and should be rejected. One could argue, especially in current days where replication is hotly debated, that the Cheng study could be seen as a welcome replication of my study. If yes, then Cheng et al. should definitely include the beta-binomial method in their simulation.

Response: We thank the reviewer for this important reference. We agree that the reviewer's recommendation based on his own simulation on using Beta-binomial regression to incorporate both-armed zero-event studies in meta-analysis is a very important addition to the literature of meta-analysis. We have added the reviewer's findings in the "Background" and "Discussion" sections (Pages 9, 30). However, before an easy-to-use statistical package made available, this rather complicated statistical model post a big challenge for most non-statisticians. Currently, most researchers use RevMan or Stata's metan procedure to conduct meta-analyses on aggregated binary outcomes. Therefore, we believe our simulation study will provide some strategies in dealing with both-armed zero-event studies for this group of researchers.

In regarding to the ethical prospective, although we recommended that both-armed zero-event studies to be excluded from the process of numerical synthesis in some situations, they will be still included in

the systematic reviews and remain as an important component of the study.

Some minor points:
- It is hard for me to accept with today's computing power that Cheng et al. only run 2.500 simulations, especially in their scenarios with only small meta-analyses. I feel 10.000 runs is a minimum here and a sensible compromise between computing time and random variation in simulation results.
Response: Based on Burton's (2006) paper (The design of simulation studies in medical statistics), we estimated the number of simulations we needed would be between 1700 and 7900 depending on between simulation variations. Although, it is not a computational burden for generating data sets, it was very time consuming to run meta-analysis using Stata metan module due to the programming limitation. Therefore, we had to compromise the number of simulations with the time.

- It was not clear for me how the true meta-analyses were generated. It felt like they were generated from a standard inverse-variance random effects model, but the authors could be more clear here.

Response: Simulated data sets were generated using R based on the simulation parameters listed in Table 1. The meta-analyses were then conducted using Stata's metan procedure. For each simulation scenario, 25,000 meta-analyses were performed (2,500 simulations X 2 approaches of including or excluding both-armed zero-events X 5 meta-analysis methods including M-H fixed and random-effects, inverse variance fixed and random-effects, and Peto). This is why running meta-analysis for each simulation scenarios took substantial time.

- It is a good tradition in methodical papers in biostatistics, that the authors illustrate their problem with an example from real data. Why did Cheng et al. not use this option?

Response: The real example of meta-analysis will have different number of patients for each study. Both-armed zero-events are more likely to be observed in smaller studies, and these small studies are normally assigned with smaller weights during the numerical synthesizing process. We worried that using the empirical data set with varied number of patients in each study may weaken our argument about "including both-armed zero-events studies may introduce bias". It could give the readers an impression that it is OK to include both-armed zero-event studies since those studies only had small weights. Therefore, we chose to use the purely imputed data with same numbers of patients within each simulation scenario to emphasize the potential danger of including both-armed zero-event studies in meta-analysis.

VERSION 2 – REVIEW

REVIEWER	Fiona Warren University of Exeter Medical School, UK
REVIEW RETURNED	29-Jun-2016

GENERAL COMMENTS	Review of resubmitted manuscript: The Impact of Including or Excluding Both-Armed Zero-Event Studies on Using Standard Meta-analysis Methods for Rare Event Outcome: A Simulation Study This resubmission is much improved. The paper is now much

	clearer and easier to read, although there are still some minor typos/inconsistencies and the expression is not always fully clear or grammatically correct. I am sure that the editorial process will address these minor issues. I would be pleased to see this paper (with minor revisions) accepted for publication in BMJ Open. Minor comment Results: Including BA0E studies P19: the authors state that including double-0 studies produced a smaller RMSE (compared with excluding double-0 studies) when OR=1 but this is not consistent with the results set out in Appendix 3a, i.e. the RMSE is larger for including double-0 studies when OR=1.When OR=0.5 the RMSE is very similar for both including and excluding double-0 studies, although slightly larger for including. This paragraph could be rewritten for clarity/accuracy.

VERSION 2 – AUTHOR RESPONSE

Responses to Fiona Warren's Comments
This resubmission is much improved. The paper is now much clearer and easier to read, although there are still some minor typos/inconsistencies and the expression is not always fully clear or grammatically correct. I am sure that the editorial process will address these minor issues. I would be pleased to see this paper (with minor revisions) accepted for publication in BMJ Open.

Response: We are very grateful for the efforts and time that the reviewer has spent on reviewing our manuscript. During this entire peer-reviewing process, her insightful, constructive and detailed comments helped us tremendously on improving the quality of our manuscript.

Minor comment
Results: Including BA0E studies
P19: the authors state that including double-0 studies produced a smaller RMSE (compared with excluding double-0 studies) when OR=1 but this is not consistent with the results set out in Appendix 3a, i.e. the RMSE is larger for including double-0 studies when OR=1.When OR=0.5 the RMSE is very similar for both including and excluding double-0 studies, although slightly larger for including. This paragraph could be rewritten for clarity/accuracy.

Response: We thank the reviewer for pointing out this inconsistency in details. The confusion is caused by the typographical error in Appendix 3a. When OR=1, the RMSEs obtained using the approach of including double-0 (BA0E) studies was around 0.13 instead of 1.13. We apologize for this error. When OR=0.5, the reviewer's statement (i.e. the RMSEs were similar between two approaches) is more accurate than our previous expression. We have made the changes in the paragraph in regard to these two points (page 19, highlighted in yellow). We also corrected the typos in Appendix 3a.