PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	The impact of heterogeneity and effect size on the estimation of the optimal information size: analysis of recently published meta-analyses.
AUTHORS	Garcia-Alamino, Josep; Bankhead, Clare; Heneghan, Carl; Pidduck, Nicola; Perera, Rafael

VERSION 1 - REVIEW

REVIEWER	Rebecca Turner MRC Biostatistics Unit, University of Cambridge, UK
REVIEW RETURNED	08-Feb-2017

GENERAL COMMENTS	Major comments: (1) The process of selecting meta-analyses for inclusion in the data set is not described clearly. There are problems with both the text of "Inclusion criteria" on p6 and with Figures 1 and 2. I've described some of the problems below: (1)(a) There's a discrepancy between the text on p6, stating that all meta-analyses are included from the selected Cochrane and non-Cochrane reviews, and the Results section (p9), stating that one meta-analysis was included from each review. (1)(b) Many systematic reviews report results from multiple treatment comparisons (reported as pairwise rather than network meta-analyses) and I couldn't find an explanation of how the authors chose which treatment comparison to use, in the text or the figures. (1)(c) Figure 1 doesn't make clear how a meta-analysis is chosen from each review. For example, the arrow leading to "exclude this outcome" near the top of the figure should link to another decision box, to explain what happens next. (1)(d) Figure 1 refers to the "first outcome of the meta-analysis" and the "main outcome of the meta-analysis". I suspect "meta-analysis" should be replaced by "treatment comparison" or "review" here, because each meta-analysis has only one outcome unless it's a multivariate meta-analysis. Also, how is "main outcome" defined? (1)(e) In Figure 2, I think "continuous outcome" should be replaced by "no binary outcomes", because most reviews report multiple outcomes rather than just one. Exclusions covered by "other reasons" should be explained in a footnote or in the text. (2) Some of the details of the analyses examining the impact of assuming different levels of heterogeneity need extra clarification. It's stated that the "heterogeneity = rep" analyses assume heterogeneity as reported in the meta-analysis. Does this refer to the value reported in a random-effects meta-analysis refitted to the data, for example if the original authors reported only a fixed-effect meta-analysis? The Rhodes et al. paper cited (reference 18) presents

	predictive distributions for heterogeneity in continuous outcome rather than binary outcome meta-analyses. I think the authors probably intended to cite a different paper by Rhodes et al. (Research Synthesis Methods 2016; 7: 346-370)? Were the predictive distributions chosen according to intervention comparison type as well as outcome type? (3) I'm confused about the analyses which regress the OIS obtained using TSA software against the OIS defined using generic sample size calculation software (Table 5, Figure 4d). My understanding is that the OIS (call this OIS_het) based on a specific I-squared value (I^2) can be calculated directly from the OIS assuming zero heterogeneity (OIS_0), using the following formula: OIS_het=OIS_0/(1-I^2), according to Wetterslev et al. (reference 21). If so, it would be very useful to provide this formula in the paper and the linear regression analyses should be left out. However, if Table 5 is left in the paper, it needs better explanation – if the second column has been obtained from the TSA software and the third has been obtained from the regression formula, why don't these columns agree as suggested by the text? Minor comments: (4) In the Abstract, the objective is not worded clearly. (5) In the fourth line of the results section of the Abstract, either "reviews" or "meta-analyses" should be deleted. (6) The results section of the Abstract doesn't explain what the numbers in brackets represent. I assume these are 95% confidence intervals for a difference in proportions, which are reported just after the percentages for two different groups. It's confusing to switch between the proportion and percentage scale in the same sentence, and would be much better to use the same scale throughout. This comment applies also to Table 3. (7) Is it useful to report the total number of events in the control groups of trials within all meta-analyses included? Currently, this is reported in the Abstract as well as the Results section. (8) In Table 2, the confidence intervals for the percentages could be omitted, as these are not very useful and make the table more cluttered. (9) Table 4 adds very little to what is reported in the text and Table 3, and could be omitted. (10) In Figure 4a, it's not explained that the y-axis is on the log scale. (11) In Figure 4c, the y-axis label is confusing and should also refer to proportion rather than %.

REVIEWER	Jørn Wetterslev Copenhagen Trial Unit, Dpt. 7812, Copenhagen University Hospital, Rigshospitalet, Blegdamsvej 2100 Copenhagen Ø, Denmark I'm a member of the task force at Copenhagen Trial Unit to develop the methodology and software for doing Trial Sequential Analysis of meta-analysis from systematic reviews
REVIEW RETURNED	28-Mar-2017

GENERAL COMMENTS	The authors should be commended for the effort to address an important question of how to choose among different possible assumptions, especially the perceived potential statistical

	heterogeneity in the meta-analysis (to come) for estimating the required information size (designated 'optimal information size' by the authors). However, I have some major concerns for the methods used and results obtained. First, the authors have chosen the phrasing optimal information size (OIS) even though the word optimal indicates that an information greater that OIS could be perceived as suboptimal which is not the case. This is especially not the case as in the random-effects model (DL) several different possible information sizes exist, corresponding to different number of included trials, despite the same choice of alfa, beta, control event rate (CER), relative risk reduction (RRR) addressed, and heterogeneity anticipated (Kulinskaya et al 2013, Wetterslev 2017). Therefore one might say that the optimal information size is the one where each of all future trials address the RRR estimated in the meta-analysis so far with 80% power and an alfa of 5% corresponding to a minimum required future number of trials (Kulinskaya et al 2013, Wetterslev 2017). However, such a cumulative information size will often be huge and unrealistically large (Kulinskaya et al 2013, Wetterslev 2017). A fair trade of will most often be to use the required information size (RIS) defined by Wetterslev et al 2009 based on then model-variance or Diversity adjusted required information size corresponding to a few trials more than the minimum required. Furthermore, it is unclear how the authors adjust the information size for heterogeneity in the meta-analyses presented in their manuscript. Second, I do not understand how the authors derive the formula for predicting the "OIS" by regressing the Log(OIS with heterogeneity) on the Log(OIS without heterogeneity) depicted in Table 5. What is x and what is y in the formula? In the text the authors write that it is Log(OIS with heterogeneity) and Log(OIS without heterogeneity) respectively, but I cannot see that this fits, even only approximately, with the actual values listed in the table 5. It would be helpful if the authors could provide the actual value of the Q3 in the table for each review as well? As this seems to be rather important for the conclusion in the manuscript that it is possible to estimate the OIS without specific software the authors should be much more transparent with how they fitted this regression equation and how they performed model control as well as presenting prediction intervals and the R-square for the regression. Also a scatter plot to document his "perfect correlation" would be appropriate. On the other hand Figure 4-d is superfluous as this correlation, which is close to perfect, between the estimation of the required information size with 0 heterogeneity in the TSA program and the Power & Sample Size is expected as the two programs use the same formula for estimating sample size. Moreover, I'm not sure that it is wise to estimate the required information size due to the associated potential heterogeneity based on a regression analysis of only 137 meta-analyses, which to a very large degree, shown by the authors themselves, don't achieve the required information size! This is because, as the cited by the authors from Thorlund et al, it may take more than 15 trials in a cumulative meta-analysis for the statistical heterogeneity to become stable and that most meta-analyses include less than 7-8 trials. Third, the authors conclude that the type of binary outcome (all cause mortality, semi-objective, and subjective) impacts the estimation of the required information size. However, due to the listed differences of heterogeneity, CER and RRR addressed, it may

	rather be these differences than the outcome type per se. Further, I'm not convinced, based on the limited number of meta-analyses included, that these differences achieve statistical significance and the authors do not give any confidence intervals (CI's) or P-values for these differences as well as they do not tell us which tests they used to demonstrate the significant differences. On the contrary we are provided with what I perceive as 95% CI's for the differences between the percentages of the included reviews meta-analyses that achieve the required information size ("OIS") in the Cochrane and the non-Cochrane meta-analyses, however this is not essential for the conclusion that "OIS" differ between meta-analyses with different types of outcomes? Fourth, it would be appropriate if the authors initially cited Turner et al: The impact of study size on meta-analyses: Examination of under-powered studies in Cochrane Reviews PLoS One volume 8, issue 3, 2013 which shows that nearly 80% of all Cochrane reviews meta-analyses do not have 80% power to detect or reject a 30% RRR, and 98% don not have 80% power to detect a 10% RRR. Therefore it is unsurprising that the authors using 5% RRR don't find any meta-analysis that achieve the RIS for a 5% RRR especially because the sample of 14 meta-analyses having all-cause mortality as outcome is very small compared to the Turner et al 2013 analysis of all Cochrane reviews at the time. Fifth, even though a cumulative meta-analysis does not reach the required information size it may still be conclusive if one of the trial sequential monitoring boundaries (for benefit, harm or futility) are crossed see Imberger et al BMJ Open 2016 and Zaina Int jr Cardiology 2013. It would be appropriate if the authors would acknowledge this as this is the main reason to address an anticipated intervention effect and the required information size in a TSA. The real problem is to investigate whether a given meta-analytic method (including a specific estimate of the RIS) gives a reliable answer to the question whether an intervention works or not with acceptable uncertainty and this can only be achieved by empirical studies (Imberger et al BMJ Open 2016) and simulation studies with transparent and realistic assumptions. Sixth, it is very confusing that the authors consistently refer to Figure 3., presenting whiskers and box-plots with medians, when writing about the mean values of CER, I-square, OIS etc. Probably it is Table 2. That they should refer to? Seventh, see the attached additional references that ought to be referenced. The reviewer also provided a file in addition to these comments. Please contact the publisher for full details.

VERSION 1 – AUTHOR RESPONSE

Reviewer:1
Reviewer Name: Rebecca Turner
Institution and Country: MRC Biostatistics Unit, University of Cambridge, UK
Competing Interests: None declared

Dear Dr Turner,

Thank you for your thoughtful and considered feedback. In response we have made the following changes:

Major comments:

1. The process of selecting meta-analyses for inclusion in the data set is not described clearly. There are problems with both the text of “Inclusion criteria” on p6 and with Figures 1 and 2. I’ve described some of the problems below:

Response: We have made changes to Figure 1 (better labelling) and Figure 2 (complete re-drawing) to clarify the selection process of how and which meta-analyses were included in our study.

(1)(a) There’s a discrepancy between the text on p6, stating that all meta-analyses are included from the selected Cochrane and non-Cochrane reviews, and the Results section (p9), stating that one meta-analysis was included from each review.

Response: We agree with the reviewer and we have modified the text accordingly introducing the following modification on p6 as follow:

“From all the selected Cochrane and non-Cochrane reviews, we included one meta-analysis from each. Based on the order the outcomes were reported (e.g. outcome 1.1 for Cochrane SRs) we selected the first outcome presented in the meta-analysis that was based on: binary data from two or more individual studies (clinical trials or randomized controlled trials). If the first outcome did not meet this inclusion criteria we continued through the listed outcomes until one was identified or we had exhausted the list of outcomes reported. (Figure 1). Meta-analyses that included observational studies, of diagnostic interventions or that were based on network meta-analysis were excluded. Meta-analyses showing no-effect (pooled effect = 1), or meta-analyses with no events in all included trials were also excluded. “

(1)(b) Many systematic reviews report results from multiple treatment comparisons (reported as pairwise rather than network meta-analyses) and I couldn’t find an explanation of how the authors chose which treatment comparison to use, in the text or the figures.

Response: We have modified the text accordingly (see above). We have also re-labelled Figure 1 as well as adding a foot note (*) to clarify this.

(1)(c) Figure 1 doesn’t make clear how a meta-analysis is chosen from each review. For example, the arrow leading to “exclude this outcome” near the top of the figure should link to another decision box, to explain what happens next.

Response: We have modified Figure 1 to be more consistent and clear.

(1)(d) Figure 1 refers to the “first outcome of the meta-analysis” and the “main outcome of the meta-

analysis". I suspect "meta-analysis" should be replaced by "treatment comparison" or "review" here, because each meta-analysis has only one outcome unless it's a multivariate meta-analysis. Also, how is "main outcome" defined?

Response: We have re-labelled Figure 1 aiming to clarify the selection process. We now refer to the selection of a single meta-analysis (main comparison) in the systematic review.

(1)(e) In Figure 2, I think "continuous outcome" should be replaced by "no binary outcomes", because most reviews report multiple outcomes rather than just one. Exclusions covered by "other reasons" should be explained in a footnote or in the text.

Response: We have changed this in agreement with the reviewer. Figure 2 has been completely re-drawn.

(2) Some of the details of the analyses examining the impact of assuming different levels of heterogeneity need extra clarification. It's stated that the "heterogeneity = rep" analyses assume heterogeneity as reported in the meta-analysis. Does this refer to the value reported in a random-effects meta-analysis refitted to the data, for example if the original authors reported only a fixed-effect meta-analysis?

Response: This value is that obtained from our re-analysis of the data using a random effects model (for consistency) even if the authors used a fixed effect model. We have added the following sentence: "heterogeneity = rep" as that reported in the meta-analysis using a random-effects model (or obtained from fitting a random effects model if a fixed effect model was used originally); (page 7).

(3) The Rhodes et al. paper cited (reference 18) presents predictive distributions for heterogeneity in continuous outcome rather than binary outcome meta-analyses. I think the authors probably intended to cite a different paper by Rhodes et al. (Research Synthesis Methods 2016; 7: 346-370)?

Response: The correct reference as stated by the reviewer is: Research Synthesis Methods 2016; 7: 346-370. We have changed this in the text.

3aWere the predictive distributions chosen according to intervention comparison type as well as outcome type?

Response:: The predictive distributions were chosen only according to outcome type and mean study size of 50-200.

(4) I'm confused about the analyses which regress the OIS obtained using TSA software against the OIS defined using generic sample size calculation software (Table 5, Figure 4d). My understanding is that the OIS (call this OIS_het) based on a specific I-squared value (I^2) can be calculated directly from the OIS assuming zero heterogeneity (OIS_0), using the following formula: OIS_het=OIS_0/(1-I^2), according to Wetterslev et al. (reference 21). If so, it would be very useful to provide this formula

in the paper and the linear regression analyses should be left out. However, if Table 5 is left in the paper, it needs better explanation – if the second column has been obtained from the TSA software and the third has been obtained from the regression formula, why don't these columns agree as suggested by the text?

Response: Many thanks for pointing this out. We have now deleted all reference to a regression model to incorporate heterogeneity to the calculation of the OIS and now have included this incorporation based on Wetterslev proposed formula in the Discussion:

"As expected, the estimation of the OIS assuming different levels of heterogeneity, and alpha values, showed a strong correlation. Although we used specialist software for the estimation of the OIS (TSA v0.9) it is possible to estimate this value using any software that allows sample size estimation if the heterogeneity level is assumed to be zero. Incorporation of heterogeneity can be done using a simple adjustment proposed by Wetterslev (19). This author proposes the use of an alternative index named the diversity (D2) statistic as opposed to the I2 factor However, there is currently no consensus on what measure of heterogeneity to adopt for the OIS (20, 21)."

Minor comments:
(5) In the Abstract, the objective is not worded clearly.

Response: We have rewritten the Abstract as per the Editor comment.

(5) In the fourth line of the results section of the Abstract, either "reviews" or "meta-analyses" should be deleted.

Response: Done.

(6) The results section of the Abstract doesn't explain what the numbers in brackets represent. I assume these are 95% confidence intervals for a difference in proportions, which are reported just after the percentages for two different groups. It's confusing to switch between the proportion and percentage scale in the same sentence, and would be much better to use the same scale throughout. This comment applies also to Table 3.

Response: The Abstract has been completely re-written.

(7) Is it useful to report the total number of events in the control groups of trials within all meta-analyses included? Currently, this is reported in the Abstract as well as the Results section.

Response: We have modified this and now present the number of participants/studies by publication type instead as this is more informative regarding the amount of evidence accumulated.

(8) In Table 2, the confidence intervals for the percentages could be omitted, as these are not very useful and make the table more cluttered.

Response: We have now deleted from Table 1 as recommended by reviewer.

(9) Table 4 adds very little to what is reported in the text and Table 3, and could be omitted.

Response: We have removed Table 4

(10) In Figure 4a, it's not explained that the y-axis is on the log scale.

Response: Done

(11) In Figure 4c, the y-axis label is confusing and should also refer to proportion rather than %.

Response: Done

Reviewer: 2
Reviewer Name: Jørn Wetterslev

Institution and Country: Copenhagen Trial Unit, Dpt. 7812, Copenhagen University Hospital, Rigshospitalet, Blegdamsvej 2100 Copenhagen Ø, Denmark

Competing Interests: I'm a member of the task force at Copenhagen Trial Unit to develop the methodology and software for doing Trial Sequential Analysis of meta-analysis from systematic reviews

The authors should be commended for the effort to address an important question of how to choose among different possible assumptions, especially the perceived potential statistical heterogeneity in the meta-analysis (to come) for estimating the required information size (designated 'optimal information size' by the authors). However, I have some major concerns for the methods used and results obtained.

Response:
We thank the reviewer for his insightful review and comments. In response we have made the following changes:

2a First, the authors have chosen the phrasing optimal information size (OIS) even though the word optimal indicates that an information greater that OIS could be perceived as suboptimal which is not the case. This is especially not the case as in the random-effects model (DL) several different possible information sizes exist, corresponding to different number of included trials, despite the same choice of alfa, beta, control event rate (CER), relative risk reduction (RRR) addressed, and heterogeneity anticipated (Kulinskaya et al 2013, Wetterslev 2017). Therefore one might say that the optimal information size is the one where each of all future trials address the RRR estimated in the

meta-analysis so far with 80% power and an alfa of 5% corresponding to a minimum required future number of trials (Kulinskaya et al 2013, Wetterslev 2017). However, such a cumulative information size will often be huge and unrealistically large (Kulinskaya et al 2013, Wetterslev 2017). A fair trade of will most often be to use the required information size (RIS) defined by Wetterslev et al 2009 based on then model-variance or Diversity adjusted required information size corresponding to a few trials more than the minimum required. Furthermore, it is unclear how the authors adjust the information size for heterogeneity in the meta-analyses presented in their manuscript.

Response:
We accept the argument described by Wetterslev. When initially proposed by Pogue and Jusuf, optimal information size calculation for a meta-analysis was defined as the minimum amount of information required for reliable conclusions. In this case the optimal information size does not indicate that information greater than the OIS could be perceived as suboptimal.

To reflect the uncertainty regarding which measure to use we have adjusted both the Introduction and the Discussion. In particular, we have added/edited (page 4):

“Other measures of information size have been proposed { 4-5 }, however the OIS involves a relatively simple calculation, which under some scenarios will underestimate the information required to define whether firm evidence has been reached to draw robust conclusions[6]. Brok et al. demonstrated, in a subset of Cochrane reviews, that many meta-analyses have false positive results due to insufficient information [3] and most meta-analysis do not have sufficient power to identify even moderate effects (7-8).”

And as a Limitation:

“There are several proposed statistics to define a “desirable sample size in terms of numbers of participants across all studies” (20).The OIS as described in this paper involves a relatively simple calculation, which if anything is likely to underestimate the information required to define whether firm evidence has been reached to draw robust conclusions [4] Therefore we used this definition of OIS as a measure to estimate what proportion of Systematic Reviews meet this minimum requirement.”

Regarding the following comment from the Reviewer: it is unclear how the authors adjust the information size for heterogeneity in the meta-analyses presented in their manuscript.

Response: In relation to the incorporation of heterogeneity to the calculation of the OIS, we used TSA software and explore the impact of different choices of I2 on this estimate. The estimates of I2 were obtained directly from the random effects meta-analysis, given a value of 0 or based in the third quartile of the predictive distributions chosen according to outcome type and mean study size (50-200 participants) as reported by Rhodes et al. Research Synthesis Methods 2016; 7:346-370). We have made changes to the document to clarify these methods.

2b. Second, I do not understand how the authors derive the formula for predicting the “OIS” by regressing the Log(OIS with heterogeneity) on the Log(OIS without heterogeneity) depicted in Table 5. What is x and what is y in the formula? In the text the authors write that it is Log(OIS with heterogeneity) and Log(OIS without heterogeneity) respectively, but I cannot see that this fits, even only approximately, with the actual values listed in the table 5. It would be helpful if the authors could provide the actual value of the Q3 in the table for each review as well? As this seems to be rather important for the conclusion in the manuscript that it is possible to estimate the OIS without specific software the authors should be much more transparent with how they fitted this regression equation

and how they performed model control as well as presenting prediction intervals and the R-square for the regression. Also a scatter plot to document his “perfect correlation” would be appropriate. On the other hand Figure 4-d is superfluous as this correlation, which is close to perfect, between the estimation of the required information size with 0 heterogeneity in the TSA program and the Power & Sample Size is expected as the two programs use the same formula for estimating sample size. Moreover, I’m not sure that it is wise to estimate the required information size due to the associated potential heterogeneity based on a regression analysis of only 137 meta-analyses, which to a very large degree, shown by the authors themselves, don’t achieve the required information size! This is because, as the cited by the authors from Thorlund et al, it may take more than 15 trials in a cumulative meta-analysis for the statistical heterogeneity to become stable and that most meta-analyses include less than 7-8 trials.

Response:
We completely agree with the reviewer (Reviewer 1 also points this out) and have deleted this from the paper. Instead we have added the following paragraph to the Discussion:

“As expected, the estimation of the OIS assuming different levels of heterogeneity, and alpha values, showed a strong correlation. Although we used specialist software for the estimation of the OIS (TSA v0.9) it is possible to estimate this value using any software that allows sample size estimation if the heterogeneity level is assumed to be zero. Incorporation of heterogeneity can be done using a simple adjustment proposed by Wetterslev (19). This author proposes the use of an alternative index named the diversity (D2) statistic as opposed to the I2 factor. However, there is currently no consensus on what measure of heterogeneity to adopt for the OIS (20, 21).”

2c. Third, the authors conclude that the type of binary outcome (all cause mortality, semi-objective, and subjective) impacts the estimation of the required information size. However, due to the listed differences of heterogeneity, CER and RRR addressed, it may rather be these differences than the outcome type per se. Further, I’m not convinced, based on the limited number of meta-analyses included, that these differences achieve statistical significance and the authors do not give any confidence intervals (CI’s) or P-values for these differences as well as they do not tell us which tests they used to demonstrate the significant differences. On the contrary we are provided with what I perceive as 95% CI’s for the differences between the percentages of the included reviews meta-analyses that achieve the required information size (“OIS”) in the Cochrane and the non-Cochrane meta-analyses, however this is not essential for the conclusion that “OIS” differ between meta-analyses with different types of outcomes?

Response: We agree with the reviewer that it is due to the differences in the other parameters that the OIS for different types of outcomes vary. Nevertheless, what we have identified is that these parameters seem to be more consistent within some outcomes and therefore the type of (binary) outcome could be seen as a simple proxy for the estimation of the relevant parameters for the CER, the RRR and the degree of heterogeneity. We have highlighted this in our paragraph related to implications for researchers and methodologists in the Discussion:

“This study has shown that the type of outcome when estimating the OIS can be used as a proxy for defining the basic parameters (CER, RRR, I2) required to perform the calculation. Systematic reviewers can use these results to calculate an OIS value for their primary outcome independently of the confidence they have on the specific parameters obtained from their review.”

2d. Fourth, it would be appropriate if the authors initially cited Turner et al: The impact of study size on meta-analyses: Examination of under-powered studies in Cochrane Reviews PLoS One volume 8, issue 3, 2013 which shows that nearly 80% of all Cochrane reviews meta-analyses do not have 80% power to detect or reject a 30% RRR, and 98% don not have 80% power to detect a 10% RRR. Therefore it is unsurprising that the authors using 5% RRR don't find any meta-analysis that achieve the RIS for a 5% RRR especially because the sample of 14 meta-analyses having all-cause mortality as outcome is very small compared to the Turner et al 2013 analysis of all Cochrane reviews at the time.

Response: Many thanks for pointing out this reference for us. We have included reference to this study in the Introduction
"Brok et al. demonstrated, in a subset of Cochrane reviews, that many meta-analyses have false positive results due to insufficient information [3] and Turner et al. showed that most meta-analysis do not have sufficient power to identify even moderate effects (7-8)."
Although not surprising it is still worth pointing out as the average effect observed for this outcome is around 5% RRR. If anything, this emphasizes the lack of evidence existing on patient relevant outcomes such as mortality.

2e. Fifth, even though a cumulative meta-analysis does not reach the required information size it may still be conclusive if one of the trial sequential monitoring boundaries (for benefit, harm or futility) are crossed see Imberger et al BMJ Open 2016 and Zaina Int jr Cardiology 2013. It would be appropriate if the authors would acknowledge this as this is the main reason to address an anticipated intervention effect and the required information size in a TSA. The real problem is to investigate whether a given meta-analytic method (including a specific estimate of the RIS) gives a reliable answer to the question whether an intervention works or not with acceptable uncertainty and this can only be achieved by empirical studies (Imberger et al BMJ Open 2016) and simulation studies with transparent and realistic assumptions.

Response: To address this issue we have added the following to our Discussion:

"We have focused exclusively on the calculation of a single threshold to define when/if a minimum level of evidence has been collected. However, retrospective analyses of meta-analytical results are more commonly used to inform prospective studies. For example, to determine the size of a new trial to answer definitively a question around efficacy. The use of Trial sequential methods has been proposed to identify early signals of effect with monitoring boundaries being defined by frequentist, semi-Bayesian and fully Bayesian methods {Higgins 2010, Watterslev 2017, Spence 2016} Although there is still considerable uncertainty about the estimates and the best method to use, empirical studies have provided examples to suggest these methods could help detect signals early (benefit, harm or futility). {Imberger 2016, Albalawi 2013}. Of note, the identification of the sample size required in a new study or studies will depend on the method used in the meta-analysis. {Kulinskaya 2014}.

2f. Sixth, it is very confusing that the authors consistently refer to Figure 3., presenting whiskers and box-plots with medians, when writing about the mean values of CER, I-square, OIS etc. Probably it is Table 2. That they should refer to?

Response: We agree with the reviewer and we have modified in the text to refer to the correct term (mean or median) and the correct Table or Figure.

2g. Seventh, see the attached additional references that ought to be referenced.

Response: Many thanks for providing us with the added references. Most of these have now been incorporated to our manuscript.

VERSION 2 – REVIEW

REVIEWER	Rebecca Turner UCL, UK
REVIEW RETURNED	26-Jun-2017

GENERAL COMMENTS	The authors have made a number of improvements to the manuscript. I am satisfied that my previous comments have been addressed.

REVIEWER	Jørn Wetterslev Copenhagen Trial Unit, Dpt. 7812, Copenhagen University Hospital Rigshospitalet, Denmark I'm a member of the task force at Copenhagen Trial Unit to develop theory and software for doing Trial Sequential Analysis which is freeware at www.tsa.dk.
REVIEW RETURNED	01-Jun-2017

GENERAL COMMENTS	The authors have addressed nearly all of my concerns in the revised version. However, I still find it peculiar to use the expression Optimal Information Size (OIS) when other adequate information sizes, even more "Optimal", pending the number of required trials, are at hand? Just because someone (Pogue & Yussuf) started using OIS many years ago it does not justify to propagate such a misleading concept. The correct expression in the light of Kulinskaya et al 2013 and Wetterslev et al 2017 should be: "The required information size (RIS) corresponding to the required number of trials". The RIS (using the D-square adjustment for heterogeneity) represents a fair trade off between each upcoming trial being well-powered and a realistic number of required future trials and randomized participants. The authors write that is still disputed whether D-square or I-sqaure should be used for heterogeneity adjustment for the RIS, this is akward as the RIS for a random-effects meta-analysis (REM) to be conclusive should use the variance in a REM for the calculation of the RIS (model variance based!). This will not be the case if one uses the I-square which actually has no defenders for calculation of a heterogeneity adjusted required information size! Please read the paper by Turner et al: "the impact of study size...." and Wetterslev et al 2009: "Estimating the required information size using Diversity in random-effects meta-analyses." This method further has the advantage that it can calculate RIS for any REM not only for the DerSimonian and Laird REM. Moreover, if the authors have used the TSA program and used the "variance based model" option for the heterogeneity adjustment they have actually used the D-square adjustment! This is not totally clear in the revised paper.

VERSION 2 – AUTHOR RESPONSE

Reviewer: 2
Reviewer Name: Jørn Wetterslev
Institution and Country: Copenhagen Trial Unit, Dpt. 7812, Copenhagen University Hospital Rigshospitalet, Denmark
Competing Interests: I'm a member of the task force at Copenhagen Trial Unit to develop theory and software for doing Trial Sequential Analysis which is freeware at www.tsa.dk.

The authors have addressed nearly all of my concerns in the revised version. However, I still find it peculiar to use the expression Optimal Information Size (OIS) when other adequate information sizes, even more "Optimal", pending the number of required trials, are at hand? Just because someone (Pogue & Yussuf) started using OIS many years ago it does not justify to propagate such a misleading concept. The correct expression in the light of Kulinskaya et al 2013 and Wetterslev et al 2017 should be: "The required information size (RIS) corresponding to the required number of trials". The RIS (using the D-square adjustment for heterogeneity) represents a fair trade off between each upcoming trial being well-powered and a realistic number of required future trials and randomized participants.

Response:
As we commented on the previous response to this question we accept the argument described by Wetterslev. However the reason to use this terminology is because it has normally been employed within the scientific community discussing these issues and will help with the dissemination of this work in the community. As an example, it has been used by GRADE as part of rating the quality of evidence:
Guyatt G, et al. GRADE guidelines 6. Rating the quality of evidence--imprecision J Clin Epidemiol. 2011 Dec;64(12):1283-93. doi: 10.1016/j.jclinepi.2011.01.012. Epub 2011 Aug 11.
Which state: “To inform this decision, one can calculate the number of patients required for an adequately powered individual trial (termed the "optimal information size" [OIS])”;
It is also referred to in Cochrane Training:
http://training.cochrane.org/resource/how-grade-evidence-imprecision

---

The authors write that is still disputed whether D-square or I-sqaure should be used for heterogeneity adjustment for the RIS, this is akward as the RIS for a random-effects meta-analysis (REM) to be conclusive should use the variance in a REM for the calculation of the RIS (model variance based!). This will not be the case if one uses the I-square which actually has no defenders for calculation of a heterogeneity adjusted required information size! Please read the paper by Turner et al: "the impact of study size...." and Wetterslev et al 2009: "Estimating the required information size using Diversity in random-effects meta-analyses." This method further has the advantage that it can calculate RIS for any REM not only for the DerSimonian and Laird REM. Moreover, if the authors have used the TSA program and used the "variance based model" option for the heterogeneity adjustment they have actually used the D-square adjustment! This is not totally clear in the revised paper.

Response:
We understand the proposal made by Wetterslev for using D-square instead of I-square and we know the papers mentioned. However the reviewer must agree that the use of I-square as an estimate of heterogeneity in a meta-analysis is currently used in the majority of published meta-analyses. Our paper aims to demystify and encourage the use of a measure of “information size” and hence the use of alternative measures (such as D-square) would reduce the utility of our findings/proposals. In our paper the more relevant point was the impact of characteristics of published Systematic Reviews on information size calculations not the comparison between I-square or D-square, which is outside the

remit of our paper.
Regarding to the TSA program we were using the mode ‘User defined’ available in the software, which allows the use of a I-square value instead of the mode ‘Model variance based’.