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)	An analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry
AUTHORS	Borah, Rohit; Brown, Andrew; Capers, Patrice; Kaiser, Kathryn

VERSION 1 - REVIEW

REVIEWER	Chrisiana Naaktgeboren, MPH PHD University Medical Center, Utrecht Netherlands
REVIEW RETURNED	20-May-2016

GENERAL COMMENTS	Thank you for the opportunitiy to read your manuscript, which as a review author, I read with interest. Your manuscript tackles the question: How long does it take to conduct a systematic review? The result was clear: a long time, something that most researchers who have ever done a review already knew. However, this review doesn't help a researcher to answer their urgent question: How long will it take to complete our proposed review? This study showed a large variation in the time needed to complete a review. For this study to be helpful, it needs to have more subgroup results. For example, how much longer do qualitative reviews take than qualitative, reviews of observational studies than that of RCTs, or reviews in which individual quality is assessed than those that don't, or reviews in which best guideline practices are followed (e.g. high score on AMSTAR) vs. those that don't. Such information will definitely be helpful for researchers planning to perform a systematic review. Additional comments Title - Change "meta-analysis" to "systematic review". Abstract – Reorder the result to put the time to complete first. Provide an IQR for this. Also, remove or rephrase this as I don't know where this conclusion comes from: "and recently published guidelines provide a framework". Methods – Please don't present results here. I think that the text box could better be incorporated in a paragraph or two on definitions. Search Efficiency – I think it would also be helpful to look into the number of final studies included compared to the number of full texts examined. In my experience, most work sits in reviewing the full text, not screening title and abstracts. Pg. 8 – Under "Reviews" – Don't include results here. Pg. 10. "Analysis of variance was used to compare means for time to complete and number of authors/team members between funded and unfunded reviews." Pg. 11. "Because of the extreme skewness for the study count

	variables from the PRISMA diagrams, we calculated z-scores and generated means, standard deviations, and ranges based on those publications with complete data that were between -2.5 and + 2.5 standard deviations in order to generate Figure 1." I am not sure what you did here. Is it simpler to say that you excluded outliers? If a study was missing number of authors, did you not, for example, include that study in the when calculating mean number of title and abstracts screened? Table 1. Please edit so that it is clearer that the 3rd and 4th columns are the IQR and the last the entire range. Table 2. I am not clear what type of analysis was done. It looks to me like only two groups are being compared to each time, but it was reported that ANOVA was used, an analysis used to compare 3 or more groups. Please clarify.

REVIEWER	Dr Andrew Booth University of Sheffield, UK
REVIEW RETURNED	04-Jun-2016

GENERAL COMMENTS	An important topic, little covered in the literature and with an interesting and innovative methodology. The study itself stands well within the acknowledged limitations of the data set. What is less forgiveable is: (1) Overclaiming the potential of metadata based on limited experience and limited empirical data (2) Misusing some of the Discussion references in order to advance their own agenda. (3) Not including in the Discussion other articles on resource use in systematic reviews. Abstract: "completion of each published" – has "review" been omitted in error or should this read "publication" instead of "published" "p<0.001" – Simply putting a p value without the values it relates to is unhelpful. The sentence should read something like "Funded reviews took significantly longer to complete and publish (Mean of X weeks vs. Y weeks, p<0.001), and involved more authors and team members (Mean A Authors/Team members vs. B Authors/Team Members, p<0.001) than those that did not report funding. The authors state in their Abstract Conclusions that: "A vision of a future in which synthesizing a group of intervention studies may be accomplished with little to no human intervention is presented". Their proposed solution does not offer this prospect as it relates to easier retrievability not to the improved synthesis and analysis stages. They also propose a transfer of effort from the endeavours of a funded systematic review team to extra overheads in the publication process – with the economic model for this labour (albeit only 15 minutes per article) and the consistency of this required skills set both being unclear. What this study adds: "conduct a systematic review and publish it," This is a cumbersome sentence construction and would read better as "conduct and then publish a systematic review" "Recent data standards proposals and informatics technology can make the process of finding and synthesizing literature much more efficient if small additions can be added to the publication process."

	This sentence requires a leap of faith particularly when using “much more efficient” for something that has not been quantified. Better as “Recent data standards proposals and informatics technology offer the potential to make the process of finding and synthesizing literature more efficient if small additions (of what?) can be added to the publication process”. “Over the last 20 years, publishing of systematic reviews has increased exponentially, as has the primary literature.” Requires reference. E.g. Bastian, H., Glasziou, P., & Chalmers, I. (2010). Seventy-five trials and eleven systematic reviews a day: how will we ever keep up?. PLoS Med,7(9), e1000326. Or something more recent. “maintained up-to-date” – this phrase is clumsy. “Per the website” – Prefer “According to the website” The authors have misused the following references in their Discussion to accentuate their argument: “A search for human clinical trials using the filters provided in PubMed with or without additional index terms provides results with unacceptably low sensitivity, precision, and specificity.14 Even something as simple as defining a study as a randomized controlled trial has been observed to be incorrect 20% of the time.15” Use of the words “unacceptably low” are neither sustained by the context – the authors have just reported yields of 3 per 100 as typical in systematic reviews and yet reference #14 reports 1 out of every 2 records as being a clinical trial, nor by the conclusions of the original authors which were “Sensitivity of the Sensitive Clinical Queries filter was reasonable (92.7%, 92.1-93.3); specificity (16.1%, 15.1-17.1) and precision were low (49.5%, 48.5-50.5).” – “unacceptably low” is a value judgement not borne out by the authors own thresholds. Similarly defining a study as an RCT being incorrect 20% of the time means that it is being described correctly 80% of the time. Constructing a straw man to advance their arguments is an “unacceptably low blow”. I was surprised that the authors have not included in their discussion more of the published estimates for time taken in reviews. The first I know of is: Allen, I. E., & Olkin, I. (1999). Estimating time to conduct a meta-analysis from number of citations retrieved. JAMA: The Journal of the American Medical Association, 282(7), 634-635. And citation searching this reference on Google Scholar will identify more recent contributions e.g.: Levay P, Ainsworth N, Kettle R, Morgan A. Identifying evidence for public health guidance: a comparison of citation searching with Web of Science and Google Scholar. Res Synth Methods. 2016 Mar;7(1):34-45. doi: 10.1002/jrsm.1158. Epub 2015 Jul 3. PubMed PMID: 26147600. Where there are some estimates for sifting times and rates. Other items that could be referred to in the Discussion include: Shemilt I, Simon A, Hollands GJ, Marteau TM, Ogilvie D, O’Mara-Eves A, et al. Pinpointing needles in giant haystacks: use of text mining to reduce impractical screening workload in extremely large scoping reviews. Res Synthesis Methods.2014;5(1):31–49. doi:

	10.1002/jrsm.1093. And the following which shows scientifically that we still have far to travel: Jonnalagadda SR, Goyal P, Huffman MD. Automating data extraction in systematic reviews: a systematic review. Systematic Reviews. 2015;4:78. doi:10.1186/s13643-015-0066-7. Included study 127 needs correction of authors “S AMCYJACSBDHSTGT”

REVIEWER	Associate Professor Tracy Merlin Managing Director, Adelaide Health Technology Assessment (AHTA), School of Public Health, University of Adelaide, South Australia Australia I have submitted systematic review protocols to the PROSPERO database.
REVIEW RETURNED	21-Jul-2016

GENERAL COMMENTS	This is a well written mansucript that tackles a topic of considerable interest to people engaged in the conduct of systematic reviews; namely, the expenditure of the effort required to collate the relevant evidence and to answer what may be a simple research or clinical question. We wade through enormous amounts of dross in order to find the occasional speck of gold! The authors have argued well that the published evidence base is only likely to increase and so there will be a concomitant increase in the need for systematic reviews. Accordingly, the current approaches to systematically reviewing the literature need to become more efficient. The research questions tested in the study are congruent with the stated objectives of the research. The authors' use of the PROSPERO database as a data source is a novel one, and entirely appropriate. However, the authors state that PROSPERO is mostly concerned with health and social care intervention studies and according to the website that is not the case - ie http://www.crd.york.ac.uk/PROSPERO/ "PROSPERO includes protocol details for systematic reviews relevant to health and social care, welfare, public health, education, crime, justice, and international development, where there is a health related outcome. Systematic review protocols on PROSPERO can include any type of any study design. Reviews of reviews and reviews of methodological issues that contain at least one outcome of direct patient or clinical relevance are also accepted." The Methods section of the manuscript is very clear and nicely lays out the study selection and data extraction methodology, as well as defining the terms used in the analysis - this proved very helpful

	when interpreting the results later in the manuscript. I have a couple of comments below regarding the methodology that may impact on the results and the conclusions that have been drawn. 1. The length of time to undertake a review was measured from the date given in the PROSPERO entry until the publication date of the review, whether sourced from PubMed or the Publishers' websites. I suspected that there would be a great degree of heterogeneity in this estimate, and this was confirmed in the results in Table 1 with a standard deviation of 31 weeks, when the mean itself was 67.3 weeks. Apart from the limitations already mentioned in the Discussion about people failing to update the registration, is the fact that systematic reviews published as reports on the local website will likely give a better overall estimate of the time taken to complete a review than those systematic reviews having to also undergo a formal peer review process and wait upon a commercial publisher to publish the review as a journal paper. The study has reported the average length of time of a completed systematic review, it has not measured the length of time to conduct a systematic review. There is a distinction. The latter may give a better estimate of the actual labour required to complete a systematic review. Perhaps both measures could be used? The latter could be estimated by estimating the time taken from entry into PROSPERO to the date of publication of the review OR the date of determining the review had not yet been recorded as completed. It is true that this latter estimate may include reviews where there was a decision not to complete or where the record was not updated but it might give a more realistic time frame for the conduct of a systematic review (particularly if you can obtain an estimate up to the point that the report is completed, as opposed to when the manuscript was published). 2. The 'authors/team members', 'time' and 'search efficiency' is likely to substantially differ according to whether the systematic review involved the review of a medical intervention/treatment, as opposed to a diagnostic/investigative service. For example, indexing of study designs are much more efficient for randomised controlled trials than for test performance (diagnostic accuracy) studies. 3. It is possible that a research team may have submitted more than one entry into PROSPERO. If this only involves two or three entries, then this is unlikely to be a problem. However, if there is significant clustering then this may have affected the estimates concerning 'authors/team members', 'funding', and potentially 'time'. 4. Figure 1 is a very effective visual depiction of the effort required to obtain a very small yield in a systematic review. It is mentioned in the text that the results of Figure 1 were skewed and the data was normalised as a consequence. Did the authors consider the reasons why the data were skewed? Could it be because there is a preference towards restriction to high level evidence? eg a systematic reviews of systematic reviews, or a systematic review of randomised controlled trials? For some review topics, this approach is not feasible, but - where feasible - this would be preferred in order to reduce the labour requirement. In the Discussion a comment is made about the inadequacy of

	MeSH indexing, and a paper by Useem et al (2015) comparing Cochrane and non-Cochrane meta-analyses is used as justification for this point. The paper by Useem et al., mentions possible reasons for the discrepancy could include sourcing English/non-English literature, different literature sources and different application of inclusion criteria. It could also include different search strategies used to search the same literature sources. I do not think this means that the use of MeSH indexing is a problem. The problem is that literature searching skills vary in those undertaking systematic reviews. For PubMed it is always best to use a search strategy that includes both MeSH and text word terms, simply because MeSH indexing can sometimes run several months to a year behind the inclusion of the paper into PubMed, but you will obtain more recent data if you search PubMed than Medline alone. I do agree with the authors that the use of meta-data, through FAIR, could help enormously in identifying research with the relevant PICOS (or PPICOS, if looking at diagnostic tests). This is certainly something to mention. There are also efforts underway to improve efficiency in the conduct of systematic reviews, through programs like RAYYAN, EpiReviewer, DistillerSR, Covidence which claim to reduce the time taken to conduct systematic reviews. There are also automatic text-mining software tools that have been trialed, albeit with limited success (eg what if a result is included in the discussion but not in the results section?; or what if a table in the results section has discrepant information from what is mentioned in the text?). I think it is over-reaching the conclusion in the abstract and the main body of the abstract that little or no human effort will be required in synthesising evidence in the future - even if the data can be obtained more efficiently, there still requires a critical appraisal to be undertaken of that evidence, a meaningful sythesis of the evidence (whether via meta-analysis or qualitatively) and interpretation of the findings. If the authors are able to address the issues raised above I think this study would be of considerable interest to systematic reviewers, commissioners of systematic reviews and readers alike.

VERSION 1 – AUTHOR RESPONSE

Reviewer: 1

Reviewer Name: Chrisiana Naaktgeboren, MPH PHD Institution and Country: University Medical Center, Utrecht, Netherlands Competing Interests: None declared

1. Thank you for the opportunity to read your manuscript, which as a review author, I read with interest. Your manuscript tackles the question: How long does it take to conduct a systematic review? The result was clear: a long time, something that most researchers who have ever done a review already knew. However, this review doesn't help a researcher to answer their urgent question: How long will it take to complete our proposed review? We found the existing data to be unsuitable for use in any type of regression/prediction equation due to the skewed distribution. We have added in some suggested references from studies where data on explicit stages of reviews may provide a better answer to this question, as well as some other support for estimates based on of the number of citations at a given stage. (14, 15)

2. This study showed a large variation in the time needed to complete a review. For this study to be helpful, it needs to have more subgroup results. For example, how much longer do qualitative reviews take than qualitative, reviews of observational studies than that of RCTs, or reviews in which individual quality is assessed than those that don't, or reviews in which best guideline practices are followed (e.g. high score on AMSTAR) vs. those that don't. Such information will definitely be helpful for researchers planning to perform a systematic review. We thank the reviewer for this suggestion, and these are certainly important questions. Since our study was unfunded and performed around other duties, we have remained focused on the stated main questions, hoping that future energy or funding may permit an analysis of the type suggested by gathering a better dataset to support these analyses. We have added notes re the limitations of this dataset to this effect in the Discussion. (14, 15)

Additional comments:

1. Title - Change "meta-analysis" to "systematic review". The editor requested clarification of the title which we have done and hope is acceptable to all concerns, and while our selection of literature to include had limited scope (as stated in the methods), we wish to avoid the implication that we undertook a traditional, exhaustive systematic review, e.g. using multiple databases as sources. We performed a meta-analysis. (1)
2. Abstract – Reorder the result to put the time to complete first. Done Provide an IQR for this. Done (2)
3. Also, remove or rephrase this as I don't know where this conclusion comes from: "and recently published guidelines provide a framework". We have changed and added additional wording to clarify this point – we hope this is acceptable. (2)
4. Methods – Please don't present results here. We present the flow of study inclusion as this is the method for determining the dataset. The results, as we interpret them, are the values we obtained from the reviews and the respective registries, not the methodological aspects of data collection itself.
5. Search Efficiency – I think it would also be helpful to look into the number of final studies included compared to the number of full texts examined. In my experience, most work sits in reviewing the full text, not screening title and abstracts. We calculated this in response to the reviewer's comment to evaluate whether this dataset might yield useful information, but the ratio of Total N Included to Full Paper Review ranged from 0 to more than 1 due to the inconsistent way in which authors report review of papers from other sources (some at title/abstract stage, some at the end stage with full paper review) and the skewness of the data prevents reliable correlations. The mean ratio of this transition to final inclusion including all data (rather than the set without the outliers in Figure 1) was .32; using the trimmed means from the figure it is .24. We have noted this in the discussion, converting to how many were *not* included by the end. (16)
6. Pg. 8 – Under "Reviews" – Don't include results here. We believe this comment is similar to #4 (please see response to #4).
7. Pg. 10. "Analysis of variance was used to compare means for time to complete and number of authors/team members between funded and unfunded reviews." We are unsure of the concern here, but if this pertains to the comment in #11, ANOVA between 2 groups is equivalent to a t-test.
8. Pg. 11. "Because of the extreme skewness for the study count variables from the PRISMA diagrams, we calculated z-scores and generated means, standard deviations, and ranges based on those publications with complete data that were between -2.5 and + 2.5 standard deviations in order to generate Figure 1." I am not sure what you did here. Is it simpler to say that you excluded outliers? We prefer to be more transparent about the exact manner in which outliers were excluded. We have rephrased this in a way that we hope is acceptable. (12)

9. If a study was missing number of authors, did you not, for example, include that study in the when calculating mean number of title and abstracts screened? We used all available data for the comparisons analyzed. All N for each analysis is reported (see tables). N.B. - No study was missing number of authors since all publications noted each author by name, rather than list a writing group, for example. (23, 24)
10. Table 1. Please edit so that it is clearer that the 3rd and 4th columns are the IQR and the last the entire range. We removed space and hope this is clearer. (23)
11. Table 2. I am not clear what type of analysis was done. It looks to me like only two groups are being compared to each time, but it was reported that ANOVA was used, an analysis used to compare 3 or more groups. Please clarify. See comment in #7 above – a t-test is a special case of ANOVA; results are the same with two groups ($F = t^2$).

Reviewer: 2

Reviewer Name: Dr Andrew Booth

Institution and Country: University of Sheffield, UK Competing Interests: None Declared

An important topic, little covered in the literature and with an interesting and innovative methodology.

The study itself stands well within the acknowledged limitations of the data set. What is less forgiveable is:

(1) Overclaiming the potential of metadata based on limited experience and limited empirical data

(2) Misusing some of the Discussion references in order to advance their own agenda.

(3) Not including in the Discussion other articles on resource use in systematic reviews.

We thank the reviewer for these thoughtful critiques and have made changes accordingly – specific instances noted below.

Abstract:

1. "completion of each published" – has "review" been omitted in error or should this read "publication" instead of "published" Corrected. (2)
2. "p<0.001" – Simply putting a p value without the values it relates to is unhelpful. The sentence should read something like "Funded reviews took significantly longer to complete and publish (Mean of X weeks vs. Y weeks, p<0.001), and involved more authors and team members (Mean A Authors/Team members vs. B Authors/Team Members, p<0.001) than those that did not report funding. We initially attempted to keep within the prescribed length limits set by the journal. We have added additional information with the consent of the editor. We hope this is acceptable. (2)
3. The authors state in their Abstract Conclusions that: "A vision of a future in which synthesizing a group of intervention studies may be accomplished with little to no human intervention is presented". Their proposed solution does not offer this prospect as it relates to easier retrievability not to the improved synthesis and analysis stages. They also propose a transfer of effort from the endeavours of a funded systematic review team to extra overheads in the publication process – with the economic model for this labour (albeit only 15 minutes per article) and the consistency of this required skills set both being unclear. We thank the reviewer for this clarifying perspective. We have revised the abstract and hope it is more

acceptable and reflective of the content of the paper as revised. We have added text to expand on these issues, with supporting references in the discussion. (2, 18)

4. What this study adds: “conduct a systematic review and publish it,” This is a cumbersome sentence construction and would read better as “conduct and then publish a systematic review” The section was removed based on note from editor. (4)
5. “Recent data standards proposals and informatics technology can make the process of finding and synthesizing literature much more efficient if small additions can be added to the publication process.” This sentence requires a leap of faith particularly when using “much more efficient” for something that has not been quantified. Better as “Recent data standards proposals and informatics technology offer the potential to make the process of finding and synthesizing literature more efficient if small additions (of what?) can be added to the publication process”. The section was removed based on note from editor. (4)
6. “Over the last 20 years, publishing of systematic reviews has increased exponentially, as has the primary literature.” Requires reference. E.g. Bastian, H., Glasziou, P., & Chalmers, I. (2010). Seventy-five trials and eleven systematic reviews a day: how will we ever keep up?. PLoS Med,7(9), e1000326. Or something more recent. So added. (6)
7. “maintained up-to-date” – this phrase is clumsy. Changed to “kept current”. (6)
8. “Per the website” – Prefer “According to the website” Changed to suggested wording. (8)
9. The authors have misused the following references in their Discussion to accentuate their argument:
    a. “A search for human clinical trials using the filters provided in PubMed with or without additional index terms provides results with unacceptably low sensitivity, precision, and specificity.14 Even something as simple as defining a study as a randomized controlled trial has been observed to be incorrect 20% of the time.15” We have changed the wording to be more conservative and expanded the point for clarity. (17)
10. Use of the words “unacceptably low” are neither sustained by the context – the authors have just reported yields of 3 per 100 as typical in systematic reviews and yet reference #14 reports 1 out of every 2 records as being a clinical trial, nor by the conclusions of the original authors which were “Sensitivity of the Sensitive Clinical Queries filter was reasonable (92.7%, 92.1-93.3); specificity (16.1%, 15.1-17.1) and precision were low (49.5%, 48.5-50.5).” – “unacceptably low” is a value judgement not borne out by the authors own thresholds. Similarly defining a study as an RCT being incorrect 20% of the time means that it is being described correctly 80% of the time. Constructing a straw man to advance their arguments is an “unacceptably low blow”. We appreciate the reviewer’s cogent arguments and have significantly restructured this section to be of what we hope is an acceptable tone. (17)
11. I was surprised that the authors have not included in their discussion more of the published estimates for time taken in reviews. The first I know of is:
    a. Allen, I. E., & Olkin, I. (1999). Estimating time to conduct a meta-analysis from number of citations retrieved. JAMA: The Journal of the American Medical Association, 282(7), 634-635.
    b. And citation searching this reference on Google Scholar will identify more recent contributions e.g.: Levay P, Ainsworth N, Kettle R, Morgan A. Identifying evidence for public health guidance: a comparison of citation searching with Web of Science and Google Scholar. Res Synth Methods. 2016 Mar;7(1):34-45. doi: 10.1002/jrsm.1158. Epub 2015 Jul 3. PubMed PMID: 26147600. Where there are some estimates for sifting times and rates.
    c. Other items that could be referred to in the Discussion include: Shemilt I, Simon A, Hollands GJ, Marteau TM, Ogilvie D, O’Mara-Eves A, et al. Pinpointing needles in giant haystacks: use of text mining to reduce impractical screening workload in extremely large scoping reviews. Res Synthesis Methods.2014;5(1):31–49. doi: 10.1002/jrsm.1093.

12. And the following which shows scientifically that we still have far to travel: Jonnalagadda SR, Goyal P, Huffman MD. Automating data extraction in systematic reviews: a systematic review. Systematic Reviews. 2015;4:78. doi:10.1186/s13643-015-0066-7.

We thank the reviewer for these suggestions, have read the papers thoroughly and incorporated them as appropriate in new text with citations. (14, 15)

13. Included study 127 needs correction of authors “S AMCYJACSBDHSTGT” Corrected, thank you. (9 of supplement)

Reviewer: 3

Reviewer Name: Associate Professor Tracy Merlin Institution and Country: Managing Director, Adelaide Health Technology Assessment (AHTA), School of Public Health, University of Adelaide, South Australia, Australia Competing Interests: I have submitted systematic review protocols to the PROSPERO database.

This is a well written manuscript that tackles a topic of considerable interest to people engaged in the conduct of systematic reviews; namely, the expenditure of the effort required to collate the relevant evidence and to answer what may be a simple research or clinical question. We wade through enormous amounts of dross in order to find the occasional speck of gold!

The authors have argued well that the published evidence base is only likely to increase and so there will be a concomitant increase in the need for systematic reviews. Accordingly, the current approaches to systematically reviewing the literature need to become more efficient.

The research questions tested in the study are congruent with the stated objectives of the research. We thank the reviewer for these kind comments.

1. The authors' use of the PROSPERO database as a data source is a novel one, and entirely appropriate. However, the authors state that PROSPERO is mostly concerned with health and social care intervention studies and according to the website that is not the case -

i.e. http://www.crd.york.ac.uk/PROSPERO/

"PROSPERO includes protocol details for systematic reviews relevant to health and social care, welfare, public health, education, crime, justice, and international development, where there is a health related outcome.

Systematic review protocols on PROSPERO can include any type of any study design. Reviews of reviews and reviews of methodological issues that contain at least one outcome of direct patient or clinical relevance are also accepted." We thank the reviewer for noting the current scope and have adjusted the text accordingly. (8)

The Methods section of the manuscript is very clear and nicely lays out the study selection and data extraction methodology, as well as defining the terms used in the analysis - this proved very helpful when interpreting the results later in the manuscript. Thank you.

I have a couple of comments below regarding the methodology that may impact on the results and the conclusions that have been drawn.

1. The length of time to undertake a review was measured from the date given in the PROSPERO entry until the publication date of the review, whether sourced from PubMed or the Publishers' websites. I suspected that there would be a great degree of heterogeneity in this estimate, and this

was confirmed in the results in Table 1 with a standard deviation of 31 weeks, when the mean itself was 67.3 weeks. Apart from the limitations already mentioned in the Discussion about people failing to update the registration, is the fact that systematic reviews published as reports on the local website will likely give a better overall estimate of the time taken to complete a review than those systematic reviews having to also undergo a formal peer review process and wait upon a commercial publisher to publish the review as a journal paper.

The study has reported the average length of time of a completed systematic review, it has not measured the length of time to conduct a systematic review. There is a distinction. The latter may give a better estimate of the actual labour required to complete a systematic review. Perhaps both measures could be used? The latter could be estimated by estimating the time taken from entry into PROSPERO to the date of publication of the review OR the date of determining the review had not yet been recorded as completed. It is true that this latter estimate may include reviews where there was a decision not to complete or where the record was not updated but it might give a more realistic time frame for the conduct of a systematic review (particularly if you can obtain an estimate up to the point that the report is completed, as opposed to when the manuscript was published). We thank the reviewer for these distinctions and certainly agree - we have added these points in the limitations section. We have also added data from other sources as suggested by another reviewer to put our results into perspective with these reports. We focused on registered reviews with publications of the results in journals, not on websites. (14, 15)

2. The 'authors/team members', 'time' and 'search efficiency' is likely to substantially differ according to whether the systematic review involved the review of a medical intervention/treatment, as opposed to a diagnostic/investigative service. For example, indexing of study designs are much more efficient for randomised controlled trials than for test performance (diagnostic accuracy) studies. We agree this could be the case, but cannot be distinguished with the data used. This point has been added as a limitation of our analysis and as a point in future research directions. (14, 15)

3. It is possible that a research team may have submitted more than one entry into PROSPERO. If this only involves two or three entries, then this is unlikely to be a problem. However, if there is significant clustering then this may have affected the estimates concerning 'authors/team members', 'funding', and potentially 'time'. We thank the reviewer for noting this possibility and indeed, this does happen, albeit rarely (3-4 instances in our dataset). We verified team member names and project titles and removed duplicate entries, retaining the most complete one. Two publications reported two separate reviews, as noted in the results section. This is noted in the methods section. (10)

4. Figure 1 is a very effective visual depiction of the effort required to obtain a very small yield in a systematic review. It is mentioned in the text that the results of Figure 1 were skewed and the data was normalised as a consequence. Did the authors consider the reasons why the data were skewed? Could it be because there is a preference towards restriction to high level evidence? eg a systematic reviews of systematic reviews, or a systematic review of randomised controlled trials? For some review topics, this approach is not feasible, but - where feasible - this would be preferred in order to reduce the labour requirement. We thank the reviewer for this observation and offer the following reply: Speculation would say that some of it is related to the specificity of the question, but also to the rigor of the group: some groups such as ours may painstakingly carry forward “maybes” at the title/abstract stage because we have found communication and reporting is too often poor or incomplete

In the Discussion a comment is made about the inadequacy of MeSH indexing, and a paper by Useem et al (2015) comparing Cochrane and non-Cochrane meta-analyses is used as justification for

this point. The paper by Useem et al., mentions possible reasons for the discrepancy could include sourcing English/non-English literature, different literature sources and different application of inclusion criteria. It could also include different search strategies used to search the same literature sources. I do not think this means that the use of MeSH indexing is a problem. The problem is that literature searching skills vary in those undertaking systematic reviews. For PubMed it is always best to use a search strategy that includes both MeSH and text word terms, simply because MeSH indexing can sometimes run several months to a year behind the inclusion of the paper into PubMed, but you will obtain more recent data if you search PubMed than Medline alone. Agreed – clarification added. (17)

I do agree with the authors that the use of meta-data, through FAIR, could help enormously in identifying research with the relevant PICOS (or PPICOS, if looking at diagnostic tests). This is certainly something to mention. Thank you.

There are also efforts underway to improve efficiency in the conduct of systematic reviews, through programs like RAYYAN, EpiReviewer, DistillerSR, Covidence which claim to reduce the time taken to conduct systematic reviews. There are also automatic text-mining software tools that have been trialed, albeit with limited success (eg what if a result is included in the discussion but not in the results section?; or what if a table in the results section has discrepant information from what is mentioned in the text?). We have added some discussion of these approaches based on 2 reviewers' comments. We did not extract data from the included studies as to whether automated procedures were used, thus potentially impacting the time to complete. This is also noted as a limitation of our data but could be a focus of future research. (15)

I think it is over-reaching the conclusion in the abstract and the main body of the abstract that little or no human effort will be required in synthesising evidence in the future - even if the data can be obtained more efficiently, there still requires a critical appraisal to be undertaken of that evidence, a meaningful synthesis of the evidence (whether via meta-analysis or qualitatively) and interpretation of the findings. We thank the reviewer for the comment and have modified our statements. (18)

If the authors are able to address the issues raised above I think this study would be of considerable interest to systematic reviewers, commissioners of systematic reviews and readers alike. Thank you. We hope we have understood the comments and made changes that are acceptable.

VERSION 2 – REVIEW

REVIEWER	Christiana Naaktgeboren UMC Utrecht, The Netherlands
REVIEW RETURNED	03-Oct-2016

GENERAL COMMENTS	Thank you for addressing my points. I have no further comments except that I am still very interested in knowing about reasons why some reviews take longer than others (e.g. diagnostic vs. therapeutic), but I will leave that up to the editor whether that is a topic for another paper.

REVIEWER	Andrew Booth School of Health & Related Research (ScHARR), University of Sheffield, UK
REVIEW RETURNED	01-Oct-2016

GENERAL COMMENTS	The Authors have worked hard and systematically to incorporate suggestions from the three reviewers and this to enhance the quality of their paper. I have considered their response and actions to the previous reviews and consider that it is now both a more complete contribution and a more informative one. Particularly placing this study in the context of other studies of resource use and opening up the prospects afforded by text mining software have enhanced the article. From a scientific perspective the authors no longer make unwarranted clams not selectively report previous articles to advance their own argumentation.

REVIEWER	A/Prof Tracy Merlin Director, Adelaide Health Technology Assessment (AHTA), School of Public Health, University of Adelaide South Australia, Australia I have submitted protocols to the PROSPERO database.
REVIEW RETURNED	11-Oct-2016

GENERAL COMMENTS	The authors have adequately addressed all of the concerns I had about the previous version of the manuscript. I have four additional minor issues that the authors are at liberty to address without affecting my recommendation to accept this version of the manuscript: 1. Systematic review registration (pg 44, line 17, tracked version) - PROSPERO accepts methodological systematic reviews if they have at least one clinical outcome. 2. Time (page 50, line 31, tracked version) - "When the article text did not include complete dates," should I think read "When the article text did not include completion dates,". 3. Last paragraph page 56, line 41, tracked version - although the reviews were matched on the same research question and were published within 5 years of each other, I think a major source of the identified differences was likely to be the different study selection criteria used in the matched reviews. This is a more likely source of the observed differences than the use of MeSH. 4. Last paragraph page 58, line 24 - "leaving only the critical appraisal to be done." should be expanded to "leaving only the critical appraisal, synthesis and interpretation to be done." As stated earlier, these suggestions involve very minor amendments and do not require further peer review.

VERSION 2 – AUTHOR RESPONSE

My fellow authors and I have made the suggested changes according to Reviewer 3's comments in what we hope is a satisfactory manner.