Reviewer reports:

Reviewer #1: General comments

This is an important question to try and answer as searching multiple databases, then downloading and de-duplicating large record sets can be very time-consuming. This paper will be of interest to information specialists and librarians involved in searching for systematic reviews, and more generally for researchers involved in systematic reviews.

Thank you.

Background/aims

1. The stated aims of paper are to determine which single or multiple database combination retrieves most relevant references. There is a second part to the study which examines current practice in recent published systematic reviews, but this is not explicitly mentioned as an aim of the paper. This section just needs a line to describe the aim of this part of the work.

We added a description of this secondary aim to our introduction
A secondary aim is to investigate the current practice of databases searched for published reviews. Are included references being missed because the review authors failed to search a certain database?

Methods

2. Development of search strategies:

It would be good to have a bit more detail here about how the search strategies are developed and then translated for different databases, especially for readers who may not be familiar with how a search strategy is developed.

The method used by the first author in creation of these search strategies is rather unique. We have submitted another article (WM Bramer, GB de Jonge, ML Rethlefsen, F Mast, J Kleijnen. A systematic approach to searching: An efficient and complete method to develop literature searches, submitted to J Med Lib Assoc) that describes in full detail the methods that we use to create the search strategies in these different databases. We have referred to the submitted articles, and have added a brief description of the method in this article.

The information specialists of Erasmus MC developed an efficient method that helps them perform searches in many databases in a much shorter time than other methods. This method of literature searching and a pragmatic evaluation thereof are published in separate journal articles [21, 22]. In short, the method consists of an efficient way to combine thesaurus terms and title/abstract terms into a single line search strategy. This search is then optimized. Articles that are indexed with a set of identified thesaurus terms, but do not contain the current search terms in title or abstract, are screened to discover potential new terms. New candidate terms are added to the basic search and evaluated. Once optimal recall is achieved, macros are used to translate the search syntaxes between databases, though manual adaptation of the thesaurus terms is still necessary.

Data analysis:

3. Did you also look at the overlap between databases, as well as the number of unique studies? What was the overall coverage of Embase and MEDLINE for example, in terms of how many included studies were retrieved by each?
Yes, this is explicated in Table 3. We believe the reviewer here means “recall” and not coverage, but if “coverage” is what is meant, we did not focus on coverage. We explicitly did not focus on coverage, as this has been done by many other studies. Most studies on this topic only look for coverage, and then conclude that since all articles are in databases X and Y, searching only that database is enough to retrieve all articles. There is an important difference between coverage and recall. The fact that an article is covered by a certain database does not mean that it is retrieved by a search strategy. This can clearly be demonstrated by the fact that the searching MEDLINE adds extra relevant articles compared to Embase (though we searched Embase without excluding MEDLINE only records). If the reviewer asks how many studies were retrieved by each, this is recall, which has been investigated in this research. This is shown in table 3 of the article.

We have studied coverage of medline embase and google scholar in another study with partially the same reviews (Bramer WM, Giustini D, Kramer BMR. Comparing the coverage, recall, and precision of searches for 120 systematic reviews in Embase, MEDLINE, and Google Scholar: a prospective study. Systematic Reviews. 2016;5(1):39). Medline alone has a coverage of 92%, Embase and Medline together cover 97%. We found that the coverage of Google Scholar is only slightly below that of Embase and Medline combined, but 3% of the articles in google scholar were only present as citations, limiting the findability of them in a search, since they could only be found by words in the title of the article.

Results
4. Of the 58 reviews analyzed, what study designs were included? In the methods you say review projects cover a wide range from effectiveness, to diagnostic, ethics etc. so what were the review types of this set? This is relevant as you comment on the lack of unique studies found in CENTRAL, which is a database of RCTs.

Based on the comments from both reviewers, we have analyzed the review types. We identified different domains, such as therapy, diagnosis, epidemiology etc. We have analyzed the difference between these groups and added that to our results. We have also included a table (Table 2) with an overview of topics and limits of the research questions in our included systematic reviews. The number of reviews which limited includes to RCTs only was only 5, while an additional set of 5 reviews included RCTs and other studies (for example cohort or case control studies). The majority did not limit the included references to study types.
Methods

To categorize the types of patient/population and intervention, we identified broad MeSH terms relating to the most important disease and intervention discussed in the article. We copied from the MeSH tree the top MeSH term directly below the disease category, or, in the case of the intervention, directly below the therapeutics MeSH term. We selected the domain from a predefined set of broad domains, including therapy, etiology, epidemiology, diagnosis, management, and prognosis. Lastly, we checked whether the reviews described limiting their included references to a certain publication type.

Differences between domains of reviews

We analyzed whether the added value of Web of Science and Google Scholar was dependent on the domain of the review. For 55 reviews, we determined the domain. See Figure 2 for the comparison of the recall of Embase, MEDLINE, and Cochrane CENTRAL per review for all identified domains. For all but one domain, the traditional combination of Embase, MEDLINE, and Cochrane CENTRAL did not retrieve enough included references. For 4 out of five systematic reviews that limited to randomized controlled trials (RCTs) only, the traditional combination retrieved 100% of all included references. However, for one review of this domain, the recall was 82%. Of the 11 references included in this review, one was found only in Google Scholar, one only in Web of Science.

5. As far as I understand, the subject specific databases were not searched for all the systematic reviews under study, but only for those where the topic was relevant, so it's not surprising that they only contribute unique references for a relevant topic. The fact that they did find unique studies is very useful to know, but we don't know either way if they would have contributed unique studies in other topics, or not.

Indeed, you are correct. We searched the subject specific databases in more reviews than only those where the topic was relevant to the topic of the database. However, we did not include these databases in all reviews that were included in this research. We have access to CINAHL and PsycINFO via our institute, but not to SportDiscus, which was included in two reviews twice via another institute. We added a paragraph on this topic to the discussion.

Of all reviews in which we searched CINAHL and PsycINFO, respectively for 6% and 9% of the reviews, unique references were found. For CINAHL and PsycINFO, in one case each, unique
relevant references were found. In both these reviews the topic was highly related to the topic of the database. Although we did not use these special topic databases in all of our reviews, given the low number of reviews where these databases added relevant references, and observing the special topics of those reviews, we suggest that these subject databases will only add value if the topic is related to the topic of the database.

6. You haven't reported the NNR in the results text. This is of interest - for example I think it's notable that the NNR of the recommended combination of databases is 73. By searching a combination of only 3 databases (ML-EM-GS) the NNR reduces to 64, with only a small drop in the overall recall.

We added a short text reference to the NNR

The higher recall from adding extra databases came at a cost in number needed to read (NNR). Searching only Embase produced an NNR of 57 on average, whereas, for the optimal combination of four databases, the NNR was 73.

Discussion

7. "The highest scoring database combination without Embase is a combination of MEDLINE, Web of Science, and Google Scholar, but that reaches satisfactory recall for only 39% of all investigated systematic reviews, while still requiring a paid subscription to Web of Science."

In this situation the addition of CENTRAL might be very beneficial, if the systematic review is focusing on RCTs, as CENTRAL contains reports of trials identified from Embase.

Thank you for this suggestion. We analyzed the added value of Cochrane CENTRAL specifically for the reviews where the search was limited to RCTs, and we added the outcome of that in the discussion of our article

Of the five reviews that included only RCTs, four reached 100% recall if MEDLINE, Web of Science, and Google Scholar combined were complemented with Cochrane CENTRAL.
8. "Subject-specific databases like PsycINFO only added unique references to a small percentage of systematic reviews in our research"

As per my previous comments, my understanding from reading the methods is that databases like this were only searched for a sub-set of the reviews?

Indeed not for all systematic reviews in our research we had used these databases. Still for 31% we used CINAHL and in 19% of our reviews we searched PsycINFO. Yet of these 18 and 11 reviews searched respectively, only one review for each database retrieved unique relevant references. In those cases the topic was clearly related to the database topic.

We clarified this in the text

Subject-specific databases like PsycINFO only added unique references to a small percentage of systematic reviews when they had been used for the search.

9. "Our research shows that the Cochrane Handbook's recommendation to search MEDLINE, Cochrane CENTRAL, and Embase is not sufficient for systematic reviews"

This is the minimum recommended for reviews of RCTs. I think your analysis included other review types.

We changed the description of the recommendation of the Cochrane Handbook accordingly and provided additional data.

The Cochrane Handbook recommends searching MEDLINE, Cochrane CENTRAL, and Embase for systematic reviews of RCTs. For reviews in our study that included RCTs only, indeed, this recommendation was sufficient for four (80%) of the reviews. The one review where it was insufficient was about alternative medicine, specifically meditation and relaxation therapy, where one of the missed studies was published in the Indian Journal of Positive Psychology. We
estimate more than 50% of reviews that include more study types than RCTs would miss more
than 5% of included references if only traditional combination of MEDLINE, Embase, and
Cochrane CENTAL is searched.

10. "To our surprise, Cochrane CENTRAL did not identify any unique references for the reviews
in this research, neither in general, nor compared with only those found in Embase and
MEDLINE, even for the three reviews focusing entirely on RCTs."

CENTRAL is a database of probable and possible RCTs. If only three of the reviews under study
focused solely on RCTs then it is not that surprising that CENTRAL did not contribute unique
studies to the other reviews, or in fact in the RCT reviews, as reports of RCTs from both PubMed
& Embase are fed into CENTRAL. I think it is worth commenting on this as at the moment it is
an unfair comparison.

We tempered our comments in the discussion by referring to this possibility.

11. "Whether a reference is available in a database is important, but whether the article can be
found in a reasonably sensitive, precise search is not only impacted by the database's coverage. It
is impacted far more by the ability of the searcher, followed closely by the accuracy of indexing
and the complexity of terminology in a particular field."

Please can you provide a reference to support this statement?

This statement is not based on research, but is a personal opinion of the authors. We changed the
second phrase to better reflect that.

Our experience has shown us that it is also impacted by the ability of the searcher, the accuracy
of indexing of the database, and the complexity of terminology in a particular field.

12. "Because these studies based on retrospective analysis of database coverage do not account
for the searchers' abilities, the actual findings from the searches performed, and the indexing for
particular articles, their conclusions are tenuous at best and misguided at worst"
I agree that a strength of the current study is its prospective design, but I think it's a bit strong to say the conclusions of previous studies are misguided, especially if the authors of those studies have acknowledged and discussed the limitations of their retrospective design.

We agree with this suggestion. We have changed the wording of this sentence.

their conclusions lack immediate translatability into practice

13. It would be helpful to know how the findings of this study compare with the previous studies - does it support or refute them?

Our conclusions are different from the earlier studies. Often it was concluded that only one database was enough to find most relevant references. No research before acknowledged the added value of searching Web of Science and Google Scholar. We have added discussion on previous findings to the discussion paragraph.

Many of the articles reporting on previous research concluded that one database was able to retrieve most included references. Halladay et al. [10] and van Enst et al. [16] concluded that databases other than MEDLINE/PubMed did not change the outcomes of the review, while Rice et al. [17] found the added value of other databases only for newer, non-indexed references. In addition, Michaleff et al. [26] found that Cochrane CENTRAL included 95% of all RCTs included in the reviews investigated. Our conclusion that Web of Science and Google Scholar are needed for completeness has not been shared by previous research. Most of the previous studies did not include these two databases in their research.

Conclusions

14. "Special topics databases such as CINAHL and PsycINFO should be added if the topic of the review is directly focused on nursing or psychiatry"

Is it possible to draw a conclusion about this as these databases were only searched for those specific topics, and not all 58 in the set?
Though we did not include these databases in every review included in our research we found that these databases only added unique included references in 6% of the reviews where we had searched the databases. And those were reviews where the topic was directly related to the focus of the database. We are confident that these databases would not had added unique relevant references to reviews of which the topic was even further away from the topic of the database. We agree that caution is required and have tempered our statements accordingly.

15. Finally - what are the authors' recommendations for future research in this topic area?

Thank you for this suggestion. We added an extra paragraph on our recommendations for future research.

Future research should continue to investigate recall of actual searches beyond coverage of databases and should consider focusing on the most optimal database combinations, not on single databases.

Reviewer #2: This is a substantial study that provides an interesting perspective, through an analysis of real-life prospective searches, on how retrieving studies for SRs is affected by the choice of database. It throws up some interesting and sometimes surprising findings. Because of this, it's incumbent on the authors to provide as full an explanation as possible. My biggest criticism is that we are not provided with more context around the differential performance of databases (and combinations of databases) as applied to individual review types (intervention, intervention RCT-only, DTAs, etc.). If there are differences according to the type, scope and criteria of reviews then this is information that will improve the usefulness and generalizability of the findings from a user perspective.

Thank you. We do agree that, ideally, the research would differentiate the effect of databases between topics and types of reviews. After acknowledging the importance of this, based on this peer review we decided to investigate in more detail the differences between review types on what we call domains: therapy, diagnosis, etiology, and epidemiology. We added that information in an extra paragraph and figure. We did not see a difference between the domains of the reviews and our most important outcome: the need to search Web of Science and Google Scholar.
Differences between domains of reviews

We analyzed whether the added value of Web of Science and Google Scholar was dependent of the domain of the review. For 55 reviews, we determined the domain. See Figure 2 for the comparison of the recall of Embase, MEDLINE, and Cochrane CENTRAL per review for all identified domains. For all but one domain, the traditional combination of Embase, MEDLINE, and Cochrane CENTRAL did not retrieve enough included references. For 4 out of five systematic reviews that limited to randomized controlled trials (RCTs) only, the traditional combination retrieved 100% of all included references. However, for one review of this domain, the recall was 82%. Of the 11 references included in this review, one was found only in Google Scholar, one only in Web of Science.

ABSTRACT

Background: The sentence containing the aim/objective should emphasize the different approach this study takes, namely assessing coverage based on prospective real-life searches as opposed to theoretical coverage.

We added an extra phrase to the Background in the Abstract

While previous studies determined coverage of databases, we analyzed the actual retrieval from the original searches for systematic reviews.

Methods (line 9/10): implies that the search results of all the information specialists at EMC are part of the study, when in fact it's only the author's searches that are included. This can be clarified in the sentence "PubMed was used..."

The reviewer is correct. All searches used in this research were done by the first of this article. We changed the first sentence of the Methods paragraph accordingly:

Since May 2013, the first author prospectively recorded results from systematic review searches that he performed at his institution.

Results (line 20/21): the finding reported in the main paper that 84 references out of the 1830 references from these 58 reviews were not identified by the SR searches is not insignificant (about 4.5% of all the references). Encourage giving consideration to incorporating this finding into the Results section.

We added that information to the first sentence of the Results section in the abstract
A total of 58 published systematic reviews were included, totaling 1746 relevant references identified by our database searches, while 84 included references had been retrieved by other search methods.

BACKGROUND

Para 1: an even bigger disadvantage to using multiple databases is having to translate thesaurus terms - I don't think this is captured in what is written now.

Thank you for this comment. We agree, and we added information about that in the first paragraph:

It is laborious for searchers to translate a search strategy into multiple interfaces and search syntaxes, as field codes and proximity operators differ between interfaces.

METHODS

Section on development of search strategies

Para 1: consider adding a footnote to explain why searches of PubMed were restricted to the publisher[sb] only. [I know this is commented on in the Discussion, but helpful for readers to know of this at this point.]

We added this into the main text instead of a footnote

PubMed restricting to records in the subset "as supplied by publisher" to find references that not yet indexed in MEDLINE (using the syntax publisher[sb]);

Para 1: you say hits from GS were restricted to 100 when number of references from other databases was low - is this correct or do you mean 1000? (Seems counter-intuitive to restrict
when other sources are low, but perhaps the logic is that if it's low in the databases, it will be low in GS.)

We did mean 100. Because we did not find many potentially relevant references in other databases, we did not expect many relevant references in Google Scholar either, therefore we limited the number of references we used from Google Scholar. We clarified this in the main text

In general, we use the first 200 references as sorted in the relevance ranking of Google Scholar. When the number of references from other databases was low, we expected the total number of potential relevant references to be low. In this case, the number of hits from Google Scholar was limited to 100.

Section on Determining relevant references of published reviews

Para 1: the word 'Retrospectively' to start the paragraph is unnecessary

We removed that word.

Section of Data analysis

- general comment: the descriptions and definitions of the outcome measures can be difficult to read and understand; it's important that terms are used unambiguously and consistently.

We have rewritten some parts. We have replaced the measure of sensitivity with recall, as this was the exact same definition and to enhance clarification.

- with so many search performance measures included in the analysis, strongly consider including a table in this section (performance measure + definition) to help with understanding and navigate through the section

We moved the definitions into a new table (Table 1). The definitions are written as formulas.

Table 1 – Definitions of general measures of performance in searches
Recall (# included references retrieved by a database/combination)/(# included references retrieved by all databases)

Precision (# included references retrieved by a database/combination)/(# total references retrieved by those database(s))

Number Needed to Read (# total references retrieved by a database/combination)/(# included references retrieved by those database(s))

- consider including a sentence in the first paragraph that summarises your primary performance measure of interest - i.e. what is it the data analysis seeking to uncover and how is this to be achieved.

We added an extra paragraph to the discussion to address this comment.

Performance of a search can be expressed in different ways. Depending on the goal of the search, different measures may be optimized. In the case of a clinical question, precision is most important, as a practicing clinician does not have a lot of time to read through many articles in a clinical setting. When searching for a systematic review, recall is the most important aspect, as the researcher does not want to miss any relevant references. As our research is performed on systematic reviews, the main performance measure is recall.

- Para 1 (last sentence): consider saying 'unique included references' rather than 'relevant' references; clarify what is meant by 'rest of the research'

Thank you for this suggestion. We changed relevant to included. The databases that contributed most unique references were considered as candidate databases for most optimal databases combinations.

Performance of a search can be expressed in different ways. Depending on the goal of the search, different measures may be optimized. In the case of a clinical question, precision is most important, as a practicing clinician does not have a lot of time to read through many articles in a clinical setting. When searching for a systematic review, recall is the most important aspect, as the researcher does not want to miss any relevant references. As our research is performed on systematic reviews, the main performance measure is recall.
- Para 2: implies that the study is calculating a single statistic to convey 'Efficiency' but I don't think this is the case - in the results overall efficiency is the equivalent to sensitivity. Please clarify.

We changed efficiency to performance throughout.

- Para 3: first sentence is a bit confusing, do you mean to say "Sensitivity was defined as the number of included references retrieved by each database (or combination of databases) divided by the total number of included references retrieved by all databases. ??

Indeed, you are correct. We changed that phrase accordingly.

- Para 4: first sentence, the word 'results' is used twice, for consistency, shouldn't this be 'references'?

We changed the second word 'results' into references.

RESULTS

- Para 1: I think it's essential to include a description of the SRs included in the sample of 58 to give added context. Were these mostly intervention effectiveness reviews? If so were these mostly RCT-only reviews? How many DTA reviews? This may give some explanation to the surprising data for CENTRAL.

Thank you for this suggestion. We added a new table 2 which contains a description of the departments of the first author, the patient group, the intervention, the domain and limits to study types. For the Patient and interventions we used broad MeSH term categories for diseases and for interventions.

Table 2 - Description of topics of included references (only values above 5% are shown)
Department (N=55)

Surgery 13 (24%)
Epidemiology 10 (18%)
Internal Medicine 3 (5%)
Orthopedics 3 (5%)

Patient (N=52)

Neoplasms 6 (12%)
Wounds and Injuries 6 (12%)
Musculoskeletal Diseases 5 (10%)
Cardiovascular Diseases 5 (10%)
Nutritional and Metabolic Diseases 5 (10%)

Intervention (N=31)

Chemicals and Drugs Category 12 (39%)
Surgical Procedures, Operative 8 (26%)
Food and Beverages 2 (6%)
Biological Factors 2 (6%)

Domain (N=54)

Therapy 19 (35%)
Etiology 13 (24%)
Epidemiology 6 (11%)
Diagnosis 6 (11%)
Management 5 (9%)
Prognosis 5 (9%)

Study Types (N=58)

No limits mentioned 48 (83%)
RCTs 5 (9%)

RCTs and cohort studies / case control studies 5 (9%)

- Para 1: consider adding "(4.6%)" after 84 references. (I know it's probably beyond the scope of this study, but that's a higher figure than I would have expected given the databases searched. Curious to know more about the sources of these missing included references.)

Thank you for this suggestion. We added the percentage to the text. Indeed we did not research these references in more detail. We are not able to research the sources ourselves, as we did not do the additional search methods ourselves. Even if we would contact each review author for information on this, it is unlikely they are able to determine retrospectively where they had identified those references.

Section on Unique references per database

- Para 1: "A total of 292" rather a 'number'

Changed

- Para 1: "... Cochrane CENTRAL, is absent from the table, as it did not add any UNIQUE included references..."

Changed

The following sections of the Results refer to the figures. As they stand at the moment, the figures are difficult to read and interpret - they require titles, ideally a legend, axis labels and an explanation included as footnote.

We have now added the captures and titles of the figures at the end of the article. We have added a new figure 3 which works as a legend to the former figures 2 and 3 (now 4 and 5).
Figure 1 - Percentage of systematic reviews for which a certain database combination reached a certain recall. The X-axis represents the percentage of reviews for which a specific combination of databases, as shown on the y-axis, reached a certain recall (represented with bar colors).

Abbreviations: EM - Embase; ML - MEDLINE; WoS - Web of Science; GS - Google Scholar

* The recall of all databases has been calculated over all included references. The recall of the database combinations was calculated over all included references retrieved by any database.

Figure 2 – Legend of Figure 3 and 4.

Figure 3 - The ratio between number of results per database combination and the total number of results for all databases

Figure 4 - The ratio between precision per database combination and the total precision for all databases

Section on Status of current practice

- I wondered if this was essential to include in the study (I'm not convinced it is.). However, I think if the implications were framed a little differently, this would help. In the results for the MEDLINE/Embase combination, you could describe more narratively: "Around a third of the reviews (37%) relied on the combination of MEDLINE and Embase. Based on our findings, this combination achieves acceptable recall about half the time (47%), implying that less than a fifth (17%) of the reviews in the PubMed sample would have achieved an acceptable recall of 95%.

This observation adds to the importance of our research. We see that two of the databases that added important unique references to our reviews are used on frequently enough. Web of Science is used in 35% and Google Scholar in 10%.

Thank you for this feedback. We changed the wording of this paragraph so that this content is hopefully clearer to readers
TABLE 1

- consider combining results column 2 and 3, e.g. 29 (50%)  

We have combined the columns 2 and 3 and 4 and 5  

- column 4 totals 293, should be 292?  

The difference was caused by articles that were not found in the four major databases included in the analyses, but was not unique as one was found in cinahl and SportDiscus and another in cinahl and PsycINFO. We have changed the data to show only the truly unique references. Therefore the number now total 291.  

- column 5 'Percentage of unique references' - this could be combined with column 4. Not clear how these data have been calculated; should the total percentage = 100 rather than 110?  

We recalculated the percentages, they now combine to 100%. The percentage is the percentage of all unique references (291) that were retrieved by a certain database. The percentages of the number of reviews is the number of reviews that found unique references divided by the number of reviews that searched a certain database.  

TABLE 2  

- column 2 #includes - consider adding (N=1746) in column header  

We added this to the header  

- it would be helpful to add footnotes, especially for median recall, minimum recall and percentage 100% recall, to aid interpretation of the data.  

We added footnotes to these headings  

Overall recall: The total number of included references retrieved by the databases divided by the total number of included references retrieved by all databases.  

Median recall: The median value of recall per review
Minimum recall: The lowest value of recall per review

Percentage 100% recall: The percentage of reviews for which the database combination retrieved all included references

Precision: The number of included references divided by the total number of references retrieved

Number Needed to Read: The total number of references retrieved divided by the number of included references

FIGURE 1
- not clear why the relative recall data for 'All databases' is worse than for the combinations of fewer databases?

In the combinations of certain databases we compared the recall of these databases against the total recall of all databases combined. This shows the effect of missing a certain database. In the case of all databases together the percentage is calculated as the fraction of the total number of included articles in the review, including the ones that were not found in the databases.

The description of figure 1 includes a caption:

* The relative recall of all databases has been calculated over all included references. The recall of the database combinations was calculated over all included references retrieved by any database.

DISCUSSION

Para 2: the opening sentence should be framed in the context of the study findings (e.g. "Our study found that....) otherwise it reads like an opinion.

That first paragraph of the discussion was indeed more an opinion than a conclusion of our study findings. We moved that paragraph to later in the discussion section. The first paragraph of the discussion does now start with: "Our study shows that…"
Para 2: second sentence: To ensure instead of To permit

Changed

Para 2: to what extent is the recommendation to search the four databases (Embase, Medline, WOS and GS) generalisable across all types of review (DTA, intervention, prognosis, etc.)? Data are collected at the individual review level, so do the stats vary according to the type of review? This seems an important point to reflect on and to inform readers about.

Thank you for this suggestion. We added commentary about generalizability to the discussion.

Para 2: PsycINFO is broader than psychiatry, perhaps say 'mental health-related reviews'

We changed this for both PsycINFO and Cinahl:

such as PsycINFO for reviews in the fields of behavioral sciences and mental health or CINAHL for reviews on the topics of nursing or allied health,

Para 3: the Cochrane Handbook recommendation refers to intervention reviews that have RCT as the inclusion criterion, not all SRs - this should be checked and clarified. It seems that a relatively small proportion of reviews were RCT-only.

We added that the recommendation by Cochrane was only "for systematic reviews of RCTs". We concluded that:

For reviews in our study that included RCTs only indeed, this turned out to be sufficient. However, this is not sufficient for systematic reviews that included other study types

Para 3: the sentence beginning "The combination of Embase and MEDLINE resulted in reasonably..." doesn't seem to fit with this paragraph.

We removed that sentence

Para 7 (final para before strengths/limitations): it says "...probability of missing more than 5% of included ... is 33% and 30% respectively." Clarify what does 'respectively' refers to here.

We clarified that this was referring to the combination Embase and Medline with either Google Scholar or Web of Science.

the probability of missing more than 5% of included references found through database searching is 33% when using Google Scholar and 30% when Web of Science is used.
Strengths and limitations

Para 1 - could also reflect on other limitations that result from the sample of reviews included in this study being broad scope and topic, especially in terms of generalisability across review types.

The only true different types of reviews we could measure were the domains. There were only a few different domains, and each had a reasonable amount of reviews. We added an extra paragraph in the limitations setting of the discussion about this:

We found that two databases previously not recommended as essential for systematic review searching, Web of Science and Google Scholar, were key to improving recall in the reviews we investigated. Because this is a novel finding, we cannot conclude whether it is due to our dataset or to a generalizable principle. It is likely that topical differences in systematic reviews may impact whether databases such as Web of Science and Google Scholar add value to the review. One explanation for our finding may be that if the research question is very specific, the topic of research might not always be mentioned in title and or abstract. In that case, Google Scholar might add value by searching the full text of articles. If the research question is more interdisciplinary, a broader science database such as Web of Science is likely to add value. The topics of the reviews studied here may simply have fallen into those categories, though the diversity of the included reviews may point to a more universal applicability.

Para 2: the last sentence reads "This separate set had retrieved..." I think this refers to the subset as supplied by the publisher, but is ambiguous because of the sentence immediately before it. Please clarify what 'separate set' refers to.

We changed the wording of that sentence. And added a short conclusion.

According to our data, PubMed’s “as supplied by publisher” subset retrieved 12 unique included references, and it was the most important addition in terms of relevant references to the four major databases. It is therefore important to search MEDLINE including the “Epub Ahead of Print, In-Process, and Other Non-Indexed Citations” references.

Para 3: second sentence, clarify what is meant by the phrase, "... and database needs..."
We changed the wording of that sentence:

The skills and experience of the searcher is one of the most important aspects in the effectiveness of systematic review search strategies.[23-25]

Para 3: sentence beginning "We do not suggest that if anyone..." the double negative confused me initially, could consider re-phrasing something like: "We are not implying that a combined search of the four recommended databases will never result in relevant references being missed, rather that failure to search any one of these four databases will lead to relevant references being missed."

We changed the sentence as suggested

Para 3: last sentence: might want to add "Our experience in this study shows that additional..."

We changed the sentence accordingly

Para 4: first sentence "... a random sample of..."

We removed the results of the random sample of reviews from PubMed.

Do the authors want to comment on what effect knowledge that these prospective searches would be later analysed for their sensitivity etc had on the first author's search performance?

This is an interesting point, however we do not believe this to be the case. Any information specialist working on systematic reviews invests as much time in the review as necessary to find all possible relevant references. That is his or her duty. The first author uses a search method that helps him to identify all relevant title abstract words and thesaurus terms to ensure no relevant references are missed. At the time of the creation of the searches this research had not been planned. Most of the registrations were done for the comparison of the coverage and recall of Embase, MEDLINE and Google Scholar (Bramer WM, Giustini D, Kramer BMR. Comparing the coverage, recall, and precision of searches for 120 systematic reviews in Embase, MEDLINE, and Google Scholar; a prospective study. Systematic Reviews. 2016;5(1):39.). Later this research was expanded to serve other goals. Reviews from the same set have also been used in another article that has recently been submitted (Bramer WM, Rethlefsen ML, Mast F, Kleijnen J. A pragmatic evaluation of a new method for librarian-mediated literature searches for
systematic reviews. submitted. 2017.). These are all intended to be part of the PhD thesis of the first author.

In the evaluation of the method for librarian mediated searches we compared (among other things) the time it took to perform the searches. The time the first author needs to perform searches for systematic reviews is much shorter than the time others need. The median time (including the reference interview and translation to all databases) for the development of the 562 systematic review searches that have until today been timed is 70 minute (compared to 12 hours for a control group). The median time for the development of the 58 reviews included in this research was 60 minutes. We therefore don't think the searches used in this research differ much from other searches by the first author.

CONCLUSIONS

- we can't verify the opening statement that "all" SRs need searches of the four databases because that information (the extent to which these results apply equally across all types of SRs is not provided to us)

We rephrased that sentence

We recommend that, regardless of their topic, searches for biomedical systematic reviews should combine Embase, MEDLINE (including electronic publications ahead of print), Web of Science (Core Collection) and Google Scholar (the 200 first relevant references) at minimum.

- PsycINFO: use of word psychiatry vs mental health

We changed this to behavioral sciences and mental health

- last sentence: consider adding to the beginning: "This study also highlights once more that searching databases alone is not enough...

We changed that accordingly
Final comment, the word probability is used to describe databases or combinations of databases missing relevant studies - I don't know if 'probability' is being in the correct statistical sense here or as a synonym for chances/likelihood, or whether this is even important.

We did not use a mathematical function to calculate probability. The definition of the Oxford Dictionary is: the extent to which an event is likely to occur, measured by the ratio of the favourable cases to the whole number of cases possible. We measured the favourable events (a database combination retrieves enough relevant references) and expressed that as a ratio of the number of observed evens (total number of reviews we observed). However, probability in this case can also be used in a less mathematical definition: The quality or state of being probable; the extent to which something is likely to happen or be the case.