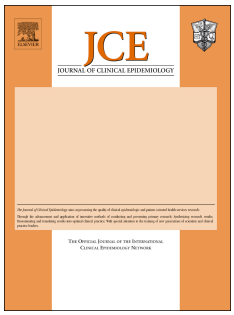
Accepted Manuscript



Systematic review identifies six metrics and one method for assessing literature search effectiveness but no consensus on appropriate use

Chris Cooper, MD, Joanna Varley-Campbell, Andrew Booth, Nicky Britten, Ruth Garside

PII: S0895-4356(17)31331-8

DOI: 10.1016/j.jclinepi.2018.02.025

Reference: JCE 9614

To appear in: *Journal of Clinical Epidemiology*

Received Date: 30 November 2017

Revised Date: 13 February 2018

Accepted Date: 27 February 2018

Please cite this article as: Cooper C, Varley-Campbell J, Booth A, Britten N, Garside R, Systematic review identifies six metrics and one method for assessing literature search effectiveness but no consensus on appropriate use, *Journal of Clinical Epidemiology* (2018), doi: 10.1016/j.jclinepi.2018.02.025.

Title:

Systematic review identifies six metrics and one method for assessing literature search effectiveness but no consensus on appropriate use

Corresponding author:

Chris Cooper, MD

Systematic Review Initiative, NHS Blood & Transplant, University of Oxford, Oxford, OX3 9BQ

Email: chris.cooper@ndcls.ox.ac.uk

Other authors:

Joanna Varley-Campbell (Centre of Outcomes Research and Effectiveness, University College London, London, WC1 7HB)
Andrew Booth (HEDS, School of Health and Related Research (ScHARR) University of Sheffield), Nicky Britten (Institute of Health Research, University of Exeter Medical School)
Ruth Garside (European Centre for Environment and Human Health, University of Exeter Medical School)

Abstract:

Objective: To identify the metrics or methods used by researchers to determine the effectiveness of literature searching where supplementary search methods are compared to bibliographic database searching. We also aimed to determine which metrics or methods are summative or formative and how researchers defined effectiveness in their studies.

Study Design and Setting: Systematic review. We searched MEDLINE and EMBASE to identify published studies evaluating literature search effectiveness in health or allied topics.

Results: Fifty studies met full-text inclusion criteria. Six metrics (Sensitivity, Specificity, Precision, Accuracy, Number Needed to Read and Yield) and one method (Capture recapture) were identified.

Conclusion: Studies evaluating effectiveness need to identify clearly the threshold at which they will define effectiveness and how the evaluation they report relates to this threshold. Studies that attempt to investigate literature search effectiveness should be informed by the reporting of confidence intervals, which aids interpretation of uncertainty around the result, and the search methods used to derive effectiveness estimates should be clearly reported and validated in studies.

Background

Various metrics or methods are used to calculate the effectiveness of literature searching. In the absence of definitive guidance, the decision on which metrics or methods can be used to evaluate literature search effectiveness is unclear. It is also unclear why researchers select the metrics they use to undertake effectiveness evaluations (1). Determining the effectiveness of literature searching can demonstrate the 'effect' of a process of literature searching, demonstrating the efficiency of a search filter, the reduction in studies to screen without missing relevant studies (time saving), and the benefits of one search approach over another.

In this systematic review, we seek to identify the metrics or methods used to calculate the effectiveness of literature searching in health and allied topics. We also seek to explore if the metrics or methods are used formatively or summatively (that is, do they seek to predict or to evaluate effectiveness (see Figure 1). This study extends beyond simply documenting how the effectiveness of literature searching has been calculated to conducting a broader examination of what effectiveness means and how it might be defined.

Methods

We followed a systematic approach to identify studies in which the calculation of literature search effectiveness was the primary objective of the study.

Research questions:
1: What metrics or methods are used to calculate literature search effectiveness?
2: Which metrics or methods are used formatively or summatively?
3: How is effectiveness defined in the studies?

Identifying studies and study data

Searching bibliographic databases

A literature search strategy was developed taking the following form: ((search terms for metrics or methods) OR (search terms for evaluation of literature searches)). This was applied to the title search field in two health-focused bibliographic databases: MEDLINE (OVID interface) and EMBASE (OVID interface). The title field was searched to identify studies in which the calculation of literature search effectiveness was the primary purpose of the study. The high prevalence of studies describing methods for literature searching, and the consequent risk of prohibitive numbers of "false hits," necessitated a strategy that placed an emphasis on search evaluation, to control the number of studies returned within resource limits for this study. Study identification was not limited by language or publication date and searches were run from database inception (MEDLINE 1946 and Embase 1974) to February 23rd 2017. The search strategies are recorded in supplementary file one.

Study selection

After visual inspection for de-duplication in Endnote X7, all studies were independently screened at title and abstract and again at full-text by two reviewers (CC and JVC).

The following inclusion criteria were applied hierarchically:

An original study published in the peer-reviewed literature that:

1. calculated literature search effectiveness;
2. provided sufficient information to replicate the calculation; and
3. calculated effectiveness between a supplementary search method (e.g. handsearching, citation chasing, web searching, contacting study authors or trials register searching) and bibliographic database searching.

The following studies were excluded:

- studies which did not compare the effectiveness of a supplementary search method against bibliographic database searching;
- studies evaluating effectiveness of teaching literature searching (i.e. trained vs. novice literature searchers);
- studies evaluating only search filters (i.e. 'search filter (a)' was compared to 'search filter (b)');
- studies evaluating the effectiveness of tools (i.e. Google Scholar vs. Web of Science); and
- abstracts, non-English language papers, letters, reviews and incomplete studies (i.e. those which do not report effectiveness outcomes).

Data extraction

Data was extracted independently into a bespoke data extraction form by CC and checked by JVC.

The following data were extracted: study citation, reference standard index test metric(s) or method(s) to calculate effectiveness, definition of effectiveness reported in the study (i.e. threshold), and claimed advantages and disadvantages relating to the calculation of effectiveness. Data were also extracted if search strategies for a reference or index test were reported and if methods to validate or quality appraise the reference standard or index test were reported. Furthermore, we determined if the evaluation was derived formatively (the purpose of the evaluation was to estimate) or summatively (the purpose of the evaluation was to calculate). The following terms are defined in figure one: reference standard, index test, summative and formative.

Quality assessment

The quality of studies was not appraised, since no appropriate quality appraisal tool exists, and this study focuses on mapping measures used and not on evaluating the studies in which they are reported.

Data synthesis

Data were synthesised narratively and summarised in tables to report the calculations for each method identified. The narrative synthesis of results was performed as follows: for each metric or method, the studies meeting full-text inclusion were read to identify the

definition of the metric or method as reported by study authors. These definitions were extracted into Microsoft excel (2013) and read repeatedly to identify commonalties or differences between definitions in the studies. A meta-definition was drafted following this exercise which was then read ('tested') against each extracted definition to ensure all the relevant aspects of definitions from the relevant studies had been captured.

Results

Database searching identified 9,126 studies for title/abstract screening after de-duplication. 200 studies were screened at full-text and 50 studies met the inclusion criteria. The Preferred Reporting in Systematic Reviews and Meta-Analysis (PRISMA) flow diagram is recorded in figure 2 (2) and studies excluded at full-text are identified in supplementary material.

Study characteristics

Of the 50 included studies (**Error! Reference source not found.**), 46 (92%) used handsearching as the reference standard. The remaining four studies used another review (n=1) or a specific combination of database searching (n=3). Validating the method or searches used to develop the reference standard was reported in 26 of 50 studies (52%) and to develop the index test in three of 50 studies (3%). Identifying a threshold to test effectiveness against was reported in 17 of 50 studies (34%). Confidence intervals were reported in 52% (26 of 50) of studies

Research Question 1 and 2: what metrics and methods are used to measure literature search effectiveness and which metrics or methods are formative or summative?

The metrics and methods used to calculate effectiveness (including specific equations) are reported in figure 3. Six metrics and one method used to calculate and evaluate literature search effectiveness were identified and had been used either individually or in combination. These metrics and methods are summarised narratively below and the calculations are reported in Table 1.

Six Metrics: summative

Sensitivity: 45/50 (90%) studies identified (3-47)

Sensitivity refers to the proportion of studies correctly identified as relevant, relative to the total number of relevant studies that may exist. All 45 studies evaluating sensitivity used the same calculation to determine a value, although the calculations are reported differently according to the type of study in which they are used (figure 3). Sensitivity is also referred to as: Recall (9, 21, 47) or relative recall[1].

[1] Eysenbach (2001) makes a distinction between recall and actual recall, as it is not truthfully possible to estimate all studies, since it is impossible to know how many unpublished studies exist at any time (48)48.

Eysenbach G, Tuische J, Diepgen TL. Evaluation of the usefulness of Internet searches to identify unpublished clinical trials for systematic reviews. Medical informatics and the Internet in medicine. 2001;26(3):203-18..

Specificity: 34/50 (68%) studies identified (4, 7, 8, 11-18, 23-26, 28-32, 34-47). Specificity refers to the number of irrelevant studies excluded or not identified by the literature search strategy. All 34 studies evaluating specificity used the same metric to determine a value (figure 3).

Precision: 40/50 (80%) studies identified (3-5, 7, 9-17, 20-26, 28-47)
Precision refers to the number of relevant studies identified by a literature search. All 40 studies used the same metric to determine a value (figure 3). Precision was also referred to as: Positive predictive value (or PPV (4, 17)).

Accuracy: 22/50 (44%) studies identified (11-16, 23-25, 30-32, 34-38, 40, 44-47) Accuracy refers to the proportion of all studies correctly identified compared to the number of non-relevant studies. All 22 studies used the same metric to determine a value (figure 3).

Number Needed to Read (NNR): 8/50 (16%) studies identified (5, 8, 9, 20, 28, 31, 41, 49). NNR is defined as the number of studies a researcher has to read to identify a relevant study. All 7 studies used the same metric to determine a value (figure 3). NNR was also referred to as: Number Needed to Search (28).

Yield (summative): 4/50 (8%) studies identified (10, 50-52)
Yield refers to the number of studies identified by a literature search method. All 4 studies interpreted yield in the same way.

Yield was often not stipulated as a metric to evaluate effectiveness but rather the yield of results from one search was directly compared with another and an assessment of effectiveness was therefore presented.

One Method: formative

Capture-Recapture (Population Estimate): 2/50 (4%) studies identified (19, 53)
Capture-Recapture (or capture mark recapture) is a formative method which provides an estimate of the 'population' of potentially relevant studies that might meet inclusion criteria.

Combinations of the above methods were commonly used. These combinations are summarised in Table 1.

Research Question 3: how is effectiveness defined in the studies?

None of the studies included in this review explicitly defined effectiveness or clearly reported what the threshold (or cut-off) was for an "effective" result in the context of their evaluation. The use of thresholds to define effectiveness were reported in 34% (17 of 50) of the studies but thresholds were commonly used to report values for inclusion of search terms into search filters (i.e. terms of min. 50% sensitivity were included), rather than as guides to interpreting the operating characteristics of the index or reference test. No study was identified that established a threshold prospectively and tested against this.

Discussion

Six metrics and one method to calculate literature search effectiveness were identified in this study. In the absence of definitive guidance, the decision on which of the metrics or methods identified in this study should be used to calculate effectiveness will continue to be determined by what researchers aim to achieve, demonstrate or explore. It is unclear how researchers selected their methods to calculate effectiveness (1).

Formative methods

Capture Re-capture was the only formative method identified and it can be used to estimate the potential number of studies to be identified from the outset of a review. This has plausible utility for allocating resources and searching time, as well as planning time to screen the number of studies identified. The Capture Re-capture method has, however, been criticised by Sampson et. al given that issues of sample independence have not been adequately explored (54).

Summative methods

The summative methods all have specific purposes when used alone: sensitivity aims to demonstrate the comprehensiveness of a literature search and NNR demonstrates the screening-rate required to identity relevant studies, for instance. When these summative methods are used in combination, researchers are able to report on effectiveness (e.g. sensitivity (55)) and efficiency (e.g. precision and NNR (55, 56))(57).

Handsearching: the 'gold standard' search method for effectiveness evaluation?

In the review, 92% of included studies used handsearching to develop their reference standard, a finding similar to a review by Jenkins (58). Handsearching aims to ensure the complete identification of studies or publication types that are not routinely indexed in, or identified by, searches of bibliographic databases, including recently published studies (59, 60). Whilst studies show that handsearching will identify studies missed by database searching (61-67), they also show that studies can be missed by handsearching (61-67), that handsearching offers low precision (61, 66) and that it is costly in terms of time (68, 69). This raises some potentially troubling questions on the suitability of handsearching as a reference standard (60, 69, 70).

Sampson et al propose an alternative to handsearching, namely the use of relative recall (68). Sampson et al define relative recall as 'the proportion that any specific system retrieves of the total or pooled relevant documents retrieved by all systems considered to be working as a composite (68).' Sampson et al's approach is a composite approach, which uses a combined set of studies as a surrogate for a reference standard and, as such, this study did not meet the inclusion criteria for this study. The disadvantages of Sampson et al's method are similar to those of handsearching: that the reference set becomes only as good as the searches that underpin it (68). Sampson et al's method would, however, mediate the concerns that calculating effectiveness using handsearching bears little relation to "real life" and it might make testing effectiveness easier, increasing the number of potential data sets available against which to test. Furthermore, since relative recall relies on underlying reviews, it might increase the transparency of methods, which would be of considerable benefit.

F Score

In peer review, a reviewer queried the absence of the F score (sometimes F-measure or F1 score) as a measure of literature search effectiveness in our review's findings. One study using F Score was identified in the main searches (71) but it did not meet inclusion at title/abstract since it did not report a calculation of literature search effectiveness between a supplementary search method and bibliographic database searching. Additional literature searches were undertaken In MEDLINE (OVID), Embase (OVID) and LISTA (EBSCOHost) to identify studies meeting our inclusion criteria and in reply to the reviewer's query. The search strategy and a PRISMA flow diagram are included in supplementary material. Thirty-nine studies were identified and double-screened. No studies met the inclusion criteria of the review.

The F Score aims to summarise precision and recall into one single number presenting a balanced mean between the two measures (72-74). As we demonstrate in this review, its application would appear to be limited in health and allied topics, and as a measure to examine literature search effectiveness. Whilst studies indicate that its use is common in information retrieval (72, 74), we found no evidence to support this.

Determining effectiveness:

Determining how effectiveness was defined in the studies was not straight-forward. We explore the issues we found, which are chiefly methodological, but this issue raises some challenging questions on the purpose of calculating effectiveness and what researchers learn by undertaking an analysis of literature search effectiveness.

Terminology:

The language used to calculate literature search effectiveness is unclear. The language used is typically borrowed from the evaluation of diagnostic tests (23) but the terms have been adopted to calculate literature search effectiveness and are used inter-changeably, often inconsistently, and sometimes confusingly between studies (14). This impairs understanding not only of what is being measured and calculated, but also what is reported and what the purpose of the calculation(s) is. Adoption of a specific and consistent language to report the calculation of literature search effectiveness would improve the transparency of effectiveness evaluation. Where possible, we have attempted to codify the language used in attempt to define the key terms relevant to the purpose of evaluating literature search effectiveness (Figure 1 and Figure 3).

Reporting and validation within studies

Whilst study quality was not formally examined, the reporting of methods to develop reference standards or index tests, and the corresponding searches undertaken, was considered poor. Only 52% of studies in the reference standard group, and 6% in the index test group, reported validating the methods and/or searches used to develop their reference standard or index test. By validation, we mean that the methods of the underlying literature search (either for the reference standard or index test) were checked or validated by another researcher. Our findings here compare with, and are arguably even worse than, those observed in a study by Patrick et al, which concluded that peer review must be developed by authors to report evidence of effectiveness of their retrieval strategies (75).

Sampson et. al have proposed a method ('Inquisitio validus Index Medicus') for search validation (54), and the Peer Review of Electronic Search Strategies (PRESS Checklist) exists for the review of electronic search methods (76). A study by Hausner et al recorded the time taken to quality appraise searches used in effectiveness evaluation as between 0.5 to 6.75 hours (77). Reporting the validation of methods used to develop reference standard or index tests, and their corresponding searches, should be a particular focus of studies seeking to calculate or estimate effectiveness of literature searching. Errors generated in producing a 'test set' will necessarily impact on the accuracy of their effectiveness estimate.

Use of thresholds

Whereas the design of studies comparing the index and reference test is self-evident, none of the studies reported a threshold beyond which they determined 'effectiveness' to have been achieved. Thirty-four percent of studies reported effectiveness thresholds (Table 2, see supplementary material), but these studies typically indicated the threshold at which search terms were included in the search strategy, rather than a prospective indication of what constituted effectiveness for the overall retrieval strategy. Gehanno et al usefully defined thresholds in their study (minimum sensitivity 65% and minimum precision 20%: NNR <5) and this approach is of benefit (9).

Diagnostic tests determine and report thresholds to indicate the point at which results are classified as either negative or positive (59). The prospective and clear reporting of thresholds in evaluation studies of search strategies would aid interpretation of the studies and would inform corresponding estimates of effectiveness generally, if the reporting of thresholds was clearer. Glanville et al prospectively determined 'ideal performance' levels for search filters through discussion with the project team. Whilst these levels were not realised within the study, their evaluation of literature search effectiveness was consequently easier to understand and analyse relative to their objectives (78).

Confidence intervals

Confidence intervals were reported in 52% of studies. Confidence intervals offer the reader an estimate of certainty (and conversely of uncertainty) in connection with the estimate of effect. Confidence intervals should, in our opinion, be calculated and reported in all studies that seek to calculate search effectiveness.

Sample size

Harbour et al reported that sample size calculations were not reported in their evaluation of search filter performance and our study shares similar conclusions (1). The number of studies included in the reference standard impacts upon the reliability of the effectiveness estimate. The reporting of sample size calculations, or alternatively why it was not considered possible to generate a reliable sample, is recommended.

Value

Effectiveness, reported in purely quantitative terms, tells researchers little about the value of the studies identified or missed, or what the effect of missing studies means (60). It is unclear what proportion of relevant studies identified represents an adequate literature search, so researchers are presently required to make their own judgements of sensitivity

(79-81). Sensitivity values do not help researchers understand this problem. It is acknowledged that no search can record 100% sensitivity (82, 83), so what does a 90% value demonstrate, other than that 10% of studies might be missing? Determining steps to identify the missing 10% (where comprehensive study identification is important to the review), or why a search was stopped, would be of benefit when reporting literature searches (84). The more pressing issue appears to be whether to revisit assumptions of the usefulness of evaluating literature searches by measuring comprehensiveness, since comprehensiveness may not be an appropriate indicator of search quality (82).

This also raises the question of what metrics or methods are most useful to record and report. Different researchers put effectiveness estimates to different purposes (5, 17, 28, 85), and it is not clear why study authors select the metrics or methods they do (1). As researchers and information specialists are being required to identify studies in new and more efficient ways, particularly in the context of abbreviated and accelerated reviews, thinking further about how effectiveness is evaluated and why, and also about what would be useful to report for other researchers, may be more important (86). Booth (2010) has called for an evaluation agenda (82). Such an agenda should be extended to include evaluating the usefulness of variables to be recorded (for instance, the time to search (38, 60) or sift is seldom recorded in studies) but it could also include different methods to capture effectiveness data (60).

Researchers may also consider how current metrics or methods may be used specifically for literature searching or making decisions on literature searching (87). A study by White et al. (published after the literature searches and screening had been completed and whilst this study was in final draft) evaluates the number needed to retrieve to justify inclusion of a database in systematic review search. This study offers 'proof of concept' testing of a metric, demonstrating that researchers can useful adapt metrics to demonstrate effectiveness, making transparent and evidence-based decisions on literature searching using data (85).

Limitations

Literature searching for this study was conducted in two bibliographic health-focused databases (MEDLINE and EMBASE). This limits the scope of this study to studies that evaluate literature search effectiveness in health or allied topics. Whilst it is a limitation in terms of scope, this limit was necessary to manage the work of the review and, methodologically, the metrics or methods identified are not limited in application to health topics. The results and discussion above apply equally to other topic areas.

This study compared effectiveness calculations between supplementary search methods and bibliographic database searching since it offered a pragmatic way to limit the scope to the resources available. The studies identified in this study are, therefore, a representative, rather than comprehensive, sample of relevant studies.

Conclusions

The review identified 50 studies that sought to calculate the effectiveness of literature searching. Whilst all 50 studies calculated the effectiveness of literature searching, what

constitutes an effective result was unclear. This leaves the question of what constitutes effectiveness in literature searching unresolved.

Studies evaluating effectiveness need to identify clearly the threshold at which they will define effectiveness and how the evaluation they report correlates to this threshold. We found that this is not yet common practice.

Studies that attempt to investigate literature search effectiveness should be informed by the reporting of confidence intervals, which aids interpretation of uncertainty within the result, and the search methods used to derive effectiveness estimates should be clearly reported and clearly validated in studies.

Acknowledgements: Juan Talens-Bou for his assistance in document ordering. Danica Cooper for proof-reading the draft manuscript.

Funding: This work forms a chapter of Chris Cooper's PhD. Chris' PhD was funded by an NIHR Health Technology Grant held at the University of Exeter.

RG and NB were partially supported by the National Institute for Health Research (NIHR) Collaboration for Leadership in Applied Health Research and Care South West Peninsula.

The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.

References

1. Harbour J, Fraser C, Lefebvre C, Glanville J, Beale S, Boachie C, et al. Reporting methodological search filter performance comparisons: a literature review. Health Information & Libraries Journal. 2014;31(3):176-94.
2. Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. Journal of Clinical Epidemiology. 2009;62:1006-12.
3. Adams CE, Power A, Frederick K, Lefebvre C. An investigation of the adequacy of MEDLINE searches for randomized controlled trials (RCTs) of the effects of mental health care. Psychological Medicine. 1994;24(3):741-8.
4. Astin MP, Brazzelli MG, Fraser CM, Counsell CE, Needham G, Grimshaw JM. Developing a sensitive search strategy in MEDLINE to retrieve studies on assessment of the diagnostic performance of imaging techniques. Radiology. 2008;247(2):365-73.
5. Bachmann LM, Coray R, Estermann P, Ter Riet G. Identifying diagnostic studies in MEDLINE: reducing the number needed to read. Journal of the American Medical Informatics Association. 2002;9(6):653-8.
6. Cathey J, Al Hajeri AA, Fedorowicz Z. A comparison of handsearching versus EMBASE searching of the Annals of Saudi Medicine to identify reports of randomized controlled trials. Annals of Saudi Medicine. 2006;26(1):49-51.
7. Dumbrigue HB, Esquivel JF, Jones JS. Assessment of MEDLINE search strategies for randomized controlled trials in prosthodontics. Journal of Prosthodontics. 2000;9(1):8-13.
8. Geersing GJ, Bouwmeester W, Zuithoff P, Spijker R, Leeflang M, Moons K. Search filters for finding prognostic and diagnostic prediction studies in medline to enhance systematic reviews. PLoS ONE. 2012;7 (2) (no pagination)(e32844).
9. Gehanno JF, Rollin L, Le Jean T, Louvel A, Darmoni S, Shaw W. Precision and recall of search strategies for identifying studies on return-to-work in Medline. Journal of Occupational Rehabilitation. 2009;19(3):223-30.
10. Glanville JM, Duffy S, McCool R, Varley D. Searching ClinicalTrials.gov and the International Clinical Trials Registry Platform to inform systematic reviews: what are the optimal search approaches? Journal of the Medical Library Association. 2014;102(3):177-83.
11. Haynes RB, Kastner M, Wilczynski NL. Developing optimal search strategies for detecting clinically sound and relevant causation studies in EMBASE. BMC Medical Informatics and Decision Making. 2005;5 (no pagination)(8).
12. Haynes RB, McKibbon KA, Wilczynski NL, Walter SD, Werre SR. Optimal search strategies for retrieving scientifically strong studies of treatment from Medline: Analytical survey. British Medical Journal. 2005;330(7501):1179-82.

13. Haynes RB, Wilczynski N, McKibbon KA, Walker CJ, Sinclair JC. Developing optimal search strategies for detecting clinically sound studies in MEDLINE. Journal of the American Medical Informatics Association. 1994;1(6):447-58.
14. Haynes RB, Wilczynski NL. Optimal search strategies for retrieving scientifically strong studies of diagnosis from Medline: analytical survey. BMJ. 2004;328(7447):1040.
15. Hildebrand AM, Iansavichus AV, Haynes RB, Wilczynski NL, Mehta RL, Parikh CR, et al. High-performance information search filters for acute kidney injury content in PubMed, Ovid Medline and Embase. Nephrology Dialysis Transplantation. 2014;29(4):823-32.
16. Holland JL, Wilczynski NL, Haynes RB. Optimal search strategies for identifying sound clinical prediction studies in EMBASE. BMC medical informatics and decision making. 2005;5:11.
17. Ingui BJ, Rogers MA. Searching for clinical prediction rules in MEDLINE. Journal of the American Medical Informatics Association : JAMIA. 2001;8(4):391-7.
18. Jenuwine ES, Floyd JA. Comparison of Medical Subject Headings and text-word searches in MEDLINE to retrieve studies on sleep in healthy individuals. Journal of the Medical Library Association. 2004;92(3):349-53.
19. Kassaï B, Sonié S, Shah NR, Boissel J-P. Literature search parameters marginally improved the pooled estimate accuracy for ultrasound in detecting deep venous thrombosis. Journal of clinical epidemiology. 2006;59(7):710-4.
20. Layton DM, Clarke M. Search Strategy to Identify Dental Survival Analysis Articles Indexed in MEDLINE. International Journal of Prosthodontics. 2016;29(1):20-7.
21. Linder SK, Kamath GR, Pratt GF, Saraykar SS, Volk RJ. Citation searches are more sensitive than keyword searches to identify studies using specific measurement instruments. Journal of Clinical Epidemiology. 2015;68(4):412-7.
22. Marson AG, Chadwick DW. How easy are randomized controlled trials in epilepsy to find on Medline? The sensitivity and precision of two Medline searches. Epilepsia. 1996;37(4):377-80.
23. McKibbon KA, Wilczynski NL, Haynes RB. Developing optimal search strategies for retrieving qualitative studies in PsycINFO. Evaluation & the Health Professions. 2006;29(4):440-54.
24. McKibbon KA, Wilczynski NL, Haynes RB. Retrieving randomized controlled trials from medline: A comparison of 38 published search filters. Health Information and Libraries Journal. 2009;26(3):187-202.
25. McKinlay RJ, Wilczynski NL, Haynes RB. Optimal search strategies for detecting cost and economic studies in EMBASE. BMC Health Services Research. 2006;6 (no pagination)(67).
26. Montori VM, Wilczynski NL, Morgan D, Haynes RB. Optimal search strategies for retrieving systematic reviews from Medline: Analytical survey. British Medical Journal. 2005;330(7482):68-71.
27. Nasser M, Al Hajeri A. A comparison of handsearching versus embase searching of the archives of Iranian medicine to identify reports of randomized controlled trials. Archives of Iranian Medicine. 2006;9(3):192-5.
28. Rogerson TE, Ladhani M, Mitchell R, Craig JC, Webster AC. Efficient strategies to find diagnostic test accuracy studies in kidney journals. Nephrology. 2015;20(8):513-8.
29. Taljaard M, McGowan J, Grimshaw JM, Brehaut JC, McRae A, Eccles MP, et al. Electronic search strategies to identify reports of cluster randomized trials in MEDLINE: low precision will improve with adherence to reporting standards. BMC Medical Research Methodology. 2010;10:15.

30. Ugolini D, Neri M, Casilli C, Bonassi S. Development of search filters for retrieval of literature on the molecular epidemiology of cancer. Mutation Research - Genetic Toxicology and Environmental Mutagenesis. 2010;701(2):107-10.
31. van de Glind EM, van Munster BC, Spijker R, Scholten RJ, Hooft L. Search filters to identify geriatric medicine in Medline. Journal of the American Medical Informatics Association. 2012;19(3):468-72.
32. Walters LA, Wilczynski NL, Haynes RB. Developing optimal search strategies for retrieving clinically relevant qualitative studies in EMBASE. Qualitative Health Research. 2006;16(1):162-8.
33. Watson RJ, Richardson PH. Identifying randomized controlled trials of cognitive therapy for depression: comparing the efficiency of Embase, Medline and PsycINFO bibliographic databases. The British journal of medical psychology. 1999;72 (Pt 4):535-42.
34. Wilczynski NL, Haynes RB. Developing optimal search strategies for detecting clinically sound causation studies in MEDLINE. Amia 2003;Annual Symposium proceedings / AMIA Symposium. AMIA Symposium.:719-23.
35. Wilczynski NL, Haynes RB. Optimal search strategies for detecting clinically sound prognostic studies in EMBASE: an analytic survey. Journal of the American Medical Informatics Association. 2005;12(4):481-5.
36. Wilczynski NL, Haynes RB. EMBASE search strategies achieved high sensitivity and specificity for retrieving methodologically sound systematic reviews. Journal of Clinical Epidemiology. 2007;60(1):29-33.
37. Wilczynski NL, Haynes RB, Eady A, Haynes B, Marks S, McKibbon A, et al. Developing optimal search strategies for detecting clinically sound prognostic studies in MEDLINE: An analytic survey. BMC Medicine. 2004;2 (no pagination)(23).
38. Wilczynski NL, Haynes RB, Hedges T. Optimal search strategies for identifying mental health content in MEDLINE: An analytic survey. Annals of General Psychiatry. 2006;5 (no pagination)(4).
39. Wilczynski NL, Haynes RB, Lavis JN, Ramkissoonsingh R, Arnold-Oatley AE. Optimal search strategies for detecting health services research studies in MEDLINE. Cmaj. 2004;171(10):1179-85.
40. Wilczynski NL, Haynes RB, Team QIH. Optimal search filters for detecting quality improvement studies in Medline. Quality & Safety in Health Care. 2010;19(6):e31.
41. Wilczynski NL, McKibbon KA, Haynes RB. Search filter precision can be improved by NOTing out irrelevant content. AMIA Annual Symposium Proceedings/AMIA Symposium. 2011;2011:1506-13.
42. Wilczynski NL, Walker CJ, McKibbon KA, Haynes RB. Assessment of methodologic search filters in MEDLINE. Proceedings - the Annual Symposium on Computer Applications in Medical Care. 1993:601-5.
43. Wilczynski NL, Walker CJ, McKibbon KA, Haynes RB. Quantitative comparison of pre-explosions and subheadings with methodologic search terms in MEDLINE. Proceedings Symposium on Computer Applications in Medical Care. 1994:905-9.
44. Wong SS, Wilczynski NL, Haynes RB. Developing optimal search strategies for detecting clinically relevant qualitative studies in MEDLINE. Medinfo. 2004;MEDINFO. 11(Pt 1):311-6.
45. Wong SS, Wilczynski NL, Haynes RB. Developing optimal search strategies for detecting clinically sound treatment studies in EMBASE. Journal of the Medical Library Association. 2006;94(1):41-7.
46. Wong SS, Wilczynski NL, Haynes RB. Optimal CINAHL search strategies for identifying therapy studies and review articles. Journal of Nursing Scholarship. 2006;38(2):194-9.

47. Wong SS, Wilczynski NL, Haynes RB, Ramkissoonsingh R. Developing optimal search strategies for detecting sound clinical prediction studies in MEDLINE. Amia 2003;Annual Symposium proceedings / AMIA Symposium. AMIA Symposium.:728-32.
48. Eysenbach G, Tuische J, Diepgen TL. Evaluation of the usefulness of Internet searches to identify unpublished clinical trials for systematic reviews. Medical informatics and the Internet in medicine. 2001;26(3):203-18.
49. Mattioli S, Farioli A, Cooke RM, Baldasseroni A, Ruotsalainen J, Placidi D, et al. Hidden effectiveness? Results of hand-searching Italian language journals for occupational health interventions. Occupational & Environmental Medicine. 2012;69(7):522-4.
50. Blanc X, Collet TH, Auer R, Iriarte P, Krause J, Legare F, et al. Retrieval of publications addressing shared decision making: an evaluation of full-text searches on medical journal websites. JMIR Research Protocols. 2015;4(2):e38.
51. Glanville J, Cikalo M, Crawford F, Dozier M, McIntosh H. Handsearching did not yield additional unique FDG-PET diagnostic test accuracy studies compared with electronic searches: a preliminary investigation. Research Synthesis Methods. 2012;3(3):202-13.
52. Hopewell S, Clarke M, Lusher A, Lefebvre C, Westby M. A comparison of handsearching versus MEDLINE searching to identify reports of randomized controlled trials. Statistics in Medicine. 2002;21(11):1625-34.
53. Spoor P, Airey M, Bennett C, Greensill J, Williams R. Use of the capture-recapture technique to evaluate the completeness of systematic literature searches. BMJ. 1996;313(7053):342-3.
54. Sampson M, McGowan J. Inquisitio validus Index Medicus: A simple method of validating MEDLINE systematic review searches. Research Synthesis Methods. 2011;2(2):103-9.
55. Kok R, Verbeek JAHM, Faber B, van Dijk FJH, Hoving JL. A search strategy to identify studies on the prognosis of work disability: a diagnostic test framework. BMJ Open. 2015;5(5).
56. Waffenschmidt S, Guddat C. Searches for randomized controlled trials of drugs in MEDLINE and EMBASE using only generic drug names compared with searches applied in current practice in systematic reviews. Res Synth Methods. 2015;6(2):188-94.
57. Lee E, Dobbins M, Decorby K, McRae L, Tirilis D, Husson H. An optimal search filter for retrieving systematic reviews and meta-analyses. BMC Med Res Methodol. 2012;12:51.
58. Jenkins M. Evaluation of methodological search filters--a review. Health Info Libr J. 2004;21(3):148-63.
59. Centre for R, Dissemination. Systematic reviews – CRD's guidance for undertaking reviews in healthcare. [Edition no.] ed. York: Centre for Reviews and Dissemination, University of York; 2009.
60. Cooper C, Booth A, Britten N, Garside R. A comparison of results of empirical studies of supplementary search techniques and recommendations in review methodology handbooks: a methodological review. Systematic Reviews. 2017;6(1):234.
61. Adams CE, Power A, Frederick K, Lefebvre C. An investigation of the adequacy of MEDLINE searches for randomized controlled trials (RCTs) of the effects of mental health care. Psychol Med. 1994;24(3):741-8.
62. Armstrong R, Jackson N, Doyle J, Waters E, Howes F. It's in your hands: the value of handsearching in conducting systematic reviews of public health interventions. Journal of Public Health. 2005;27(4):388-91.
63. Croft AM, Vassallo DJ, Rowe M. Handsearching the Journal of the Royal Army Medical Corps for trials. Journal of the Royal Army Medical Corps. 1999;145(2):86-8.

64. Hay PJ, Adams CE, Lefebvre C. The efficiency of searches for randomized controlled trials in the International Journal of Eating Disorders: a comparison of handsearching, EMBASE and PsycLIT. Health Libraries Review. 1996;13(2):91-6.
65. Hopewell S, Clarke M, Lusher A, Lefebvre C, Westby M. A comparison of handsearching versus MEDLINE searching to identify reports of randomized controlled trials. Stat Med. 2002;21(11):1625-34.
66. Jadad AR, McQuay HJ. A high-yield strategy to identify randomized controlled trials for systematic reviews. The Online journal of current clinical trials. 1993;Doc No 33:[3973 words; 39 paragraphs].
67. Langham J, Thompson E, Rowan K. Identification of randomized controlled trials from the emergency medicine literature: comparison of hand searching versus MEDLINE searching. Ann Emerg Med. 1999;34(1):25-34.
68. Sampson M, Zhang L, Morrison A, Barrowman NJ, Clifford TJ, Platt RW, et al. An alternative to the hand searching gold standard: validating methodological search filters using relative recall. BMC Med Res Methodol. 2006;6:33.
69. Doust JA, Pietrzak E, Sanders S, Glasziou PP. Identifying studies for systematic reviews of diagnostic tests was difficult due to the poor sensitivity and precision of methodologic filters and the lack of information in the abstract. Journal of Clinical Epidemiology. 2005;58(5):444-9.
70. Ugolini D, Neri M, Casilli C, Bonassi S. Development of search filters for retrieval of literature on the molecular epidemiology of cancer. Mutation research. 2010;701(2):107-10.
71. Vlachos A, Craven M. Biomedical event extraction from abstracts and full papers using search-based structured prediction. BMC Bioinformatics. 2012;13(Suppl 11):S5-S.
72. Guns R, Lioma C, Larsen B. The tipping point: F-score as a function of the number of retrieved items. Information Processing & Management. 2012;48(6):1171-80.
73. Hripcsak G, Rothschild AS. Agreement, the F-Measure, and Reliability in Information Retrieval. Journal of the American Medical Informatics Association. 2005;12(3):296-8.
74. Liu Z, Tan M, Jiang F. Regularized F-measure maximization for feature selection and classification. Journal of Biomedicine & Biotechnology. 2009;2009:617946.
75. Patrick TB, Demiris G, Folk LC, Moxley DE, Mitchell JA, Tao D. Evidence-based retrieval in evidence-based medicine. Journal of the Medical Library Association. 2004;92(2):196-9.
76. McGowan J, Sampson M, Salzwedel DM, Cogo E, Foerster V, Lefebvre C. PRESS Peer Review of Electronic Search Strategies: 2015 Guideline Statement. Journal of Clinical Epidemiology. 2016;75:40-6.
77. Hausner E, Guddat C, Hermanns T, Lampert U, Waffenschmidt S. Development of search strategies for systematic reviews: validation showed the noninferiority of the objective approach. J Clin Epidemiol. 2015;68(2):191-9.
78. Glanville J, Kaunelis D, Mensinkai S. How well do search filters perform in identifying economic evaluations in MEDLINE and EMBASE. Int J Technol Assess Health Care. 2009;25(4):522-9.
79. Lane D, Dykeman J, Ferri M, Goldsmith CH, Stelfox HT. Capture-mark-recapture as a tool for estimating the number of articles available for systematic reviews in critical care medicine. Journal of Critical Care. 2013;28(4):469-75.
80. Kastner M, Straus SE, McKibbon KA, Goldsmith CH. The capture-mark-recapture technique can be used as a stopping rule when searching in systematic reviews. J Clin Epidemiol. 2009;62(2):149-57.
81. Siva S. Optimal strategies for literature search. Indian Journal of Urology : IJU : Journal of the Urological Society of India. 2009;25(2):246-50.

82. Booth A. How much searching is enough? Comprehensive versus optimal retrieval for technology assessments. Int J Technol Assess Health Care. 2010;26(4):431-5.
83. Hildebrand AM, Iansavichus AV, Lee CW, Haynes RB, Wilczynski NL, McKibbon KA, et al. Glomerular disease search filters for Pubmed, Ovid Medline, and Embase: a development and validation study. BMC Med Inform Decis Mak. 2012;12:49.
84. Webster AJ, Kemp R. Estimating Omissions From Searches. The American Statistician. 2013;67(2):82-9.
85. Ross-White A, Godfrey C. Is there an optimum number needed to retrieve to justify inclusion of a database in a systematic review search? Health Info Libr J. 2017.
86. Egan M, MacLean A, Sweeting H, Hunt K. Comparing the effectiveness of using generic and specific search terms in electronic databases to identify health outcomes for a systematic review: a prospective comparative study of literature search methods. BMJ Open. 2012;2(3).
87. Booth A. The number needed to retrieve: a practically useful measure of information retrieval? Health Info Libr J. 2006;23(3):229-32.

1 Table 1. Included studies

2

		Reference Standard	Index	Metric or Method						
				Sensitivity	Specificity	Precision	Accuracy	NNR	Yield	Other
1	Adams 1994	Hand searching	Identify RCTs in MEDLINE	X		X				
2	Astin 2008	Hand searching	Identify diagnostic studies in MEDLINE	X	X	X				
3	Bachmann 2002	Hand searching	Identify diagnostic studies in MEDLINE	X		X		X		
4	Blanc 2015	PubMed searches	Web-searching						X	
5	Cathey 2006	Hand searching	Searching EMBASE	X						
6	Dumbridge 2000	Hand searching	Identify RCTs in MEDLINE	X	X	X				
7	Geersing 2012	Hand searching	Identify prognostic and diagnostic studies in MEDLINE	X	X			X		
8	Gehanno 2009	Hand searching and PubMed searches	Identify studies in MEDLINE	X		X		X		
9	Glanville 2012	Hand searching	Electronic searching for diagnostic studies						X	Time
10	Glanville 2014	Searching CDSR/ MEDLINE	Searching trials registries	X		X			X	
11	Haynes 1994	Hand searching	Identify studies in MEDLINE	X	X	X	X			
12	Haynes 2004	Hand searching	Identify studies in MEDLINE	X	X	X	X			
13	Haynes 2005[a]	Hand searching	Identify studies in MEDLINE	X	X	X	X			
14	Haynes 2005[b]	Hand searching	Identify studies in EMBASE	X	X	X	X			
15	Hilderbrand 2014	Hand searching	Identify studies in PubMed, MEDLINE and EMBASE	X	X	X	X			
16	Holland 2005	Hand searching	Identify studies in EMBASE	X	X	X	X			
17	Hopewell 2002	Hand searching	Identify RCTs in MEDLINE						X	
18	Ingui 2001	Hand searching and other reviews	Identify studies in MEDLINE	X	X	x				
19	Jenuwine 2004	Hand searching	Identify studies in PubMed	X	X					
20	Kassai 2006	Hand searching	Identify diagnostic studies	X						Capture-recapture
21	Layton 1988	Hand searching	Identify studies in MEDLINE	X		X		X		
22	Linder 2015	Database searching	Citation chasing	X		X				
23	Marson 1996	Hand searching	Identify RCTs in MEDLINE	X		X				
24	Mattoli 2012	Hand searching	Identify studies in PubMed					X		
25	McKibbon 2006	Hand searching	Indentidy studies in PsycINFO	X	X	X	X			
26	McKibbon 2009	Hand searching	38 RCT search filters	X	X	X	X			
27	McKinlay 2006	Hand searching	Identify cost and economic studies in EMBASE	X	X	X	X			
28	Montori 2004	Hand searching	Identify SRs in MEDLINE	X	X	X				
29	Nasser 2006	Hand searching	Identify RCTs	X						
30	Rogerson 2015	Hand searching	Identify diagnostic studies	X	X	X		X		
31	Spoor 1996	Hand searching	Identify RCTs in MEDLINE							Capture recapture
32	Taljaard 2010	Hand searching	Identify RCTs in MEDLINE	X	X	X				
33	Ugolini 2010	Hand searching	Identify studies in MEDLINE	X	X	X	X			

		Reference Standard	Index	Metric or Method						
34	van de Glind 2012	Hand searching	Identify studies in MEDLINE	X	X	X	X	X		
35	Walters 2005	Hand searching	Identification of qualitative studies in EMBASE	X	X	X	X			
36	Watson 1999	Hand searching	Identify studies in MEDLINE and PsycINFO	X		X				
37	Wilczynski 1993	Hand searching	Identify studies in MEDLINE	X	X	X				
38	Wilczynski 1994	Hand searching	Identify studies in MEDLINE	X	X	X				
39	Wilczynski 2003	Hand searching	Identify studies in MEDLINE	X	X	X	X			
40	Wilczynski 2004	Hand searching	Identify prognostic studies in MEDLINE	X	X	X	X			
41	Wilczynski-2004	Hand searching	Identify studies in MEDLINE	X	X	X				
42	Wilczynski 2005	Hand searching	Identify studies in EMBASE	X	X	X	X			
43	Wilczynski 2006	Hand searching	Identify studies in MEDLINE	X	X	X	X			
44	Wilczynski 2007	Hand searching	Identify SRs in EMBASE	X	X	X	X			
45	Wilczynski 2010	Hand searching	Identify studies in MEDLINE	X	X	X	X			
46	Wilczynski 2011	Hand searching	Effect of NOT'ing content from search strategies	X	X	X		X		
47	Wong 2003	Hand searching	Identify studies in MEDLINE	X	X	X	X			
48	Wong 2004	Hand searching	Identify studies in MEDLINE	X	X	X	X			
49	Wong 2006	Hand searching	Identify studies in CINAHL	X	X	X	X			
50	Wong 2006	Hand searching	Identify studies in EMBASE	X	X	X	X			
	TOTAL			**45**	**34**	**40**	**22**	**8**	**4**	**3**

Figure 1 key terminology defined

Reference standard (s): The reference standard is usually the best test currently available and it is the standard against which the index test is compared*.

Index test: The test which is being evaluated*.

Formative: A formative method or metric provides researchers with a potential estimate of literature search effectiveness whilst the process of literature searching is on-going. An example would be estimating the likely number of potentially relevant studies that a literature search might identify.

Summative: A summative method or metric provides the researcher with data on the performance of a completed literature search. This helps to determine the effectiveness of a completed literature searching since values can only be determined when searching is completed. An example would be calculating the Number Needed to Read. This shows how many studies a researcher read to identify an includable study.

* source: Centre for Reviews and Dissemination. Systematic reviews – CRD's guidance for undertaking reviews in healthcare. York: Centre for Reviews and Dissemination, University of York; 2009.

Figure 1 PRISMA Flow Diagram

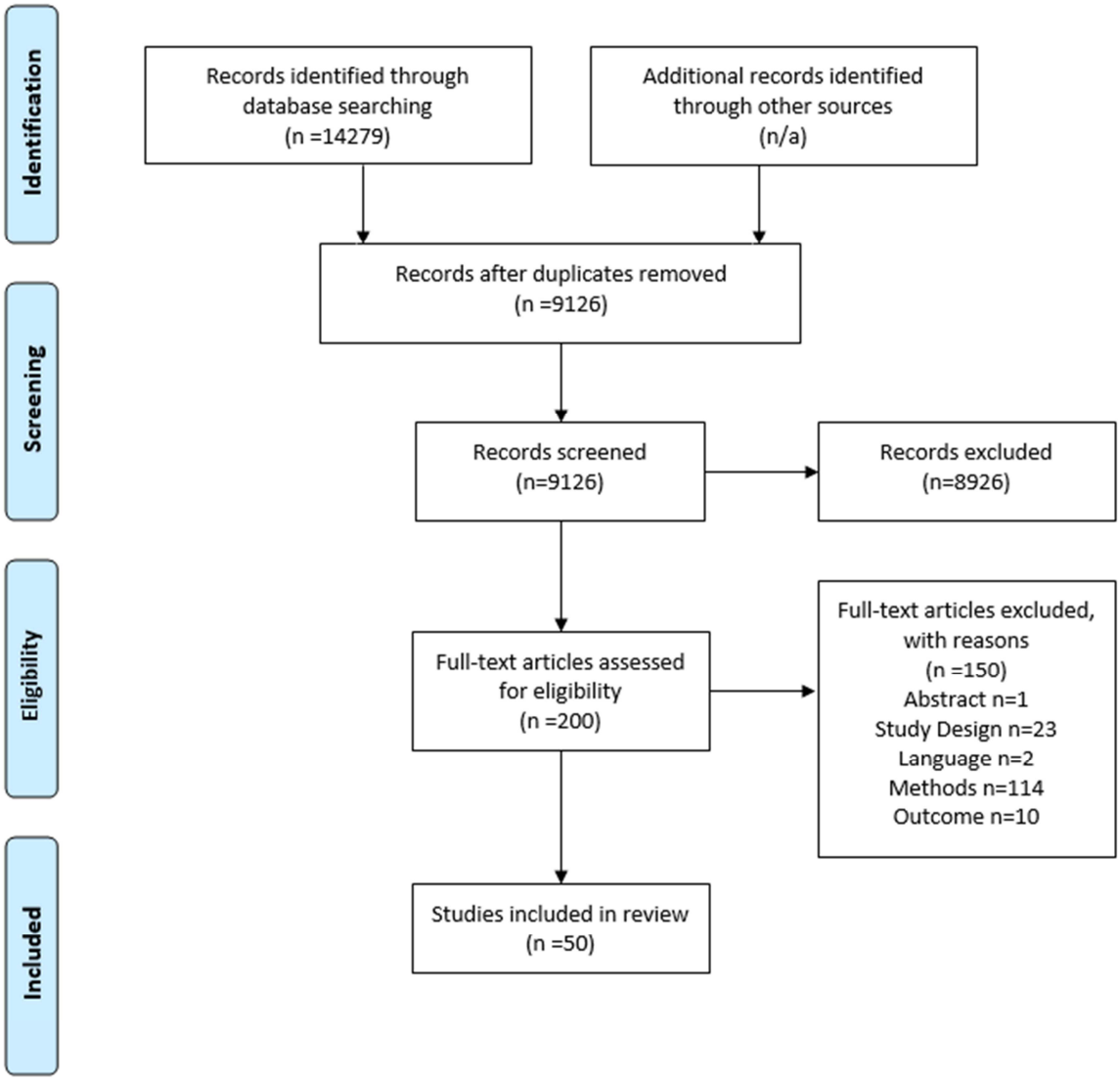


Figure 1 schematic of key metrics and methods to evaluate literature search effectiveness and their respective calculations

	Reference Standard	
Index Test	Article meets criteria (relevant)	Article does not meet criteria (not relevant)
Articles identified	a (true positives)	b (false positives)
Articles not identified	c (false negatives)	d (true negatives)

Sensitivity	The proportion of studies correctly identified as relevant, relative to the total number of relevant studies that may exist	$\frac{\text{eligible articles retrieved}}{\text{total number of eligible articles}}$ X100	a / (a + c)
Specificity	The number of irrelevant studies excluded or not identified by the literature search strategy	$\frac{\text{ineligible articles retrieved}}{\text{total number of ineligible articles}}$ X100	d / (b + d)
Precision	The number of relevant studies identified by a literature search	$\frac{\text{eligible articles retrieved}}{\text{total number of articles retrieved}}$ X100	a / (a + b)
Accuracy	The proportion of all studies correctly identified compared to the number of non-relevant studies	$\frac{\text{total number of articles retrieved}}{\text{all articles}}$ X100	(a + b) / (a + b + c + d)
Number Needed to Read (NNR)	The number of studies a researcher must read to identify a relevant study	$\frac{1}{\text{precision}}$	(a + b) / a
Yield	The number of studies identified by a literature search method	total number of articles retrieved	a + b
Capture recapture/ Population Estimate	Provides an estimate of the 'population' of potentially relevant studies that might meet inclusion criteria	number of article by search method A X $\frac{\text{number of articles by search method B}}{\text{number of articles by search method A+B}}$	x (y/z)

What's New:

Key findings: Six metrics and one method were identified that researchers have used to evaluate literature search effectiveness in health or allied topics.

What this adds to what is known: the first systematic identification and evaluation of metrics or methods to evaluate literature search effectiveness.

What is the implication, what should change now:

Studies evaluating effectiveness need to:

- identify clearly the threshold at which they will define effectiveness and how the evaluation they report relates to this threshold;
- report confidence intervals to aid the interpretation of uncertainty around the result; and
- clearly report and validate the literature search strategies used to derive effectiveness estimates.

On behalf of all of the authors:

'Declarations of interest: none'