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


This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.
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
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 commonalities 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 recall1.

---

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)

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.
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 identify 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-screens. 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


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.


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].
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.


<table>
<thead>
<tr>
<th></th>
<th>Reference Standard</th>
<th>Index</th>
<th>Metric or Method</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Precision</th>
<th>Accuracy</th>
<th>NNR</th>
<th>Yield</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Adams 1994</td>
<td>Hand searching</td>
<td>Identify RCTs in MEDLINE</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Astin 2008</td>
<td>Hand searching</td>
<td>Identify diagnostic studies in MEDLINE</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Bachmann 2002</td>
<td>Hand searching</td>
<td>Identify diagnostic studies in MEDLINE</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Blanc 2015</td>
<td>PubMed searches</td>
<td>Web-searching</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Cathey 2006</td>
<td>Hand searching</td>
<td>Searching EMBASE</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Dumbridge 2000</td>
<td>Hand searching</td>
<td>Identify RCTs in MEDLINE</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Geersing 2012</td>
<td>Hand searching</td>
<td>Identify prognostic and diagnostic studies in MEDLINE</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Gehanno 2009</td>
<td>Hand searching and PubMed searches</td>
<td>Identify studies in MEDLINE</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Glanville 2012</td>
<td>Hand searching</td>
<td>Electronic searching for diagnostic studies</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Glanville 2014</td>
<td>Searching CDSR/MEDLINE</td>
<td>Searching trials registries</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Haynes 1994</td>
<td>Hand searching</td>
<td>Identify studies in MEDLINE</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>Haynes 2004</td>
<td>Hand searching</td>
<td>Identify studies in MEDLINE</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>Haynes 2005</td>
<td>Hand searching</td>
<td>Identify studies in MEDLINE</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>Haynes 2005’</td>
<td>Hand searching</td>
<td>Identify studies in EMBASE</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>Hilderbrand 2014</td>
<td>Hand searching</td>
<td>Identify studies in PubMed, MEDLINE and EMBASE</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>Holland 2005</td>
<td>Hand searching</td>
<td>Identify studies in EMBASE</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>Hopewell 2002</td>
<td>Hand searching</td>
<td>Identify RCTs in MEDLINE</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>Ingui 2001</td>
<td>Hand searching and other reviews</td>
<td>Identify studies in MEDLINE</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>Jenuwine 2004</td>
<td>Hand searching</td>
<td>Identify studies in PubMed</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>Kassai 2006</td>
<td>Hand searching</td>
<td>Identify diagnostic studies</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Capture-recapture</td>
</tr>
<tr>
<td>21</td>
<td>Layton 1988</td>
<td>Hand searching</td>
<td>Identify studies in MEDLINE</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>Linder 2015</td>
<td>Database searching</td>
<td>Citation chasing</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>Marson 1996</td>
<td>Hand searching</td>
<td>Identify RCTs in MEDLINE</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>Mattoli 2012</td>
<td>Hand searching</td>
<td>Identify studies in PubMed</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>Mckibbon 2006</td>
<td>Hand searching</td>
<td>Identify studies in PsycINFO</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>Mckibbon 2009</td>
<td>Hand searching</td>
<td>38 RCT search filters</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>McKinlay 2006</td>
<td>Hand searching</td>
<td>Identify cost and economic studies in EMBASE</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>Montori 2004</td>
<td>Hand searching</td>
<td>Identify SRs in MEDLINE</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>Nasser 2006</td>
<td>Hand searching</td>
<td>Identify RCTs</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>Rogerson 2015</td>
<td>Hand searching</td>
<td>Identify diagnostic studies</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Capture-recapture</td>
</tr>
<tr>
<td>31</td>
<td>Spoor 1996</td>
<td>Hand searching</td>
<td>Identify RCTs in MEDLINE</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>Taljaard 2010</td>
<td>Hand searching</td>
<td>Identify RCTs in MEDLINE</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>Ugolini 2010</td>
<td>Hand searching</td>
<td>Identify studies in MEDLINE</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reference Standard</td>
<td>Index</td>
<td>Metric or Method</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------------------</td>
<td>-------</td>
<td>------------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>van de Glind 2012</td>
<td>Hand searching</td>
<td>Identify studies in MEDLINE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Walters 2005</td>
<td>Hand searching</td>
<td>Identification of qualitative studies in EMBASE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Watson 1999</td>
<td>Hand searching</td>
<td>Identify studies in MEDLINE and PsycINFO</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wilczynski 1993</td>
<td>Hand searching</td>
<td>Identify studies in MEDLINE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wilczynski 1994</td>
<td>Hand searching</td>
<td>Identify studies in MEDLINE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wilczynski 2003</td>
<td>Hand searching</td>
<td>Identify studies in MEDLINE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wilczynski 2004</td>
<td>Hand searching</td>
<td>Identify prognostic studies in MEDLINE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wilczynski 2004</td>
<td>Hand searching</td>
<td>Identify studies in MEDLINE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wilczynski 2005</td>
<td>Hand searching</td>
<td>Identify studies in EMBASE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wilczynski 2006</td>
<td>Hand searching</td>
<td>Identify studies in MEDLINE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wilczynski 2007</td>
<td>Hand searching</td>
<td>Identify SRs in EMBASE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wilczynski 2010</td>
<td>Hand searching</td>
<td>Identify studies in MEDLINE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wilczynski 2011</td>
<td>Hand searching</td>
<td>Effect of NOT'ing content from search strategies</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wong 2003</td>
<td>Hand searching</td>
<td>Identify studies in MEDLINE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wong 2004</td>
<td>Hand searching</td>
<td>Identify studies in MEDLINE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wong 2006</td>
<td>Hand searching</td>
<td>Identify studies in CINAHL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wong 2006</td>
<td>Hand searching</td>
<td>Identify studies in EMBASE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td></td>
<td>45 34 40 22 8 4 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
**Figure 1** key terminology defined

<table>
<thead>
<tr>
<th>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*.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Index test: The test which is being evaluated*.</td>
</tr>
<tr>
<td>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.</td>
</tr>
<tr>
<td>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.</td>
</tr>
</tbody>
</table>

* 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

<table>
<thead>
<tr>
<th>Index Test</th>
<th>Reference Standard</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Article meets criteria (relevant) Article does not meet criteria (not relevant)</td>
</tr>
<tr>
<td>Articles identified</td>
<td>a (true positives)</td>
</tr>
<tr>
<td>Articles not identified</td>
<td>b (false positives)</td>
</tr>
<tr>
<td>c (false negatives)</td>
<td>d (true negatives)</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>The proportion of studies correctly identified as relevant, relative to the total number of relevant studies that may exist</td>
</tr>
<tr>
<td>Specificity</td>
<td>The number of irrelevant studies excluded or not identified by the literature search strategy</td>
</tr>
<tr>
<td>Precision</td>
<td>The number of relevant studies identified by a literature search</td>
</tr>
<tr>
<td>Accuracy</td>
<td>The proportion of all studies correctly identified compared to the number of non-relevant studies</td>
</tr>
<tr>
<td>Number Needed to Read (NRR)</td>
<td>The number of studies a researcher must read to identify a relevant study</td>
</tr>
<tr>
<td>Yield</td>
<td>The number of studies identified by a literature search method</td>
</tr>
<tr>
<td>Capture recapture/Population Estimate</td>
<td>Provides an estimate of the 'population' of potentially relevant studies that might meet inclusion criteria</td>
</tr>
</tbody>
</table>

\[
\text{Sensitivity} = \frac{a}{a + c} \\
\text{Specificity} = \frac{d}{b + d} \\
\text{Precision} = \frac{a}{a + b} \\
\text{Accuracy} = \frac{(a + b)}{(a + b + c + d)} \\
\text{Number Needed to Read (NRR)} = \frac{1}{\text{precision}} \\
\text{Yield} = \frac{a + b}{\text{total number of articles retrieved}} \\
\text{Capture recapture/Population Estimate} = \frac{\text{number of article by search method A} \times \text{number of articles by search method B}}{\text{number of articles by search method A+B}}
\]
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'