



This is a repository copy of *Systematic review identifies six metrics and one method for assessing literature search effectiveness but no consensus on appropriate use*.

White Rose Research Online URL for this paper:
<http://eprints.whiterose.ac.uk/128645/>

Version: Accepted Version

Article:

Cooper, C., Varley-Campbell, J., Booth, A. orcid.org/0000-0003-4808-3880 et al. (2 more authors) (2018) Systematic review identifies six metrics and one method for assessing literature search effectiveness but no consensus on appropriate use. *Journal of Clinical Epidemiology*, 99. pp. 53-63. ISSN 0895-4356

<https://doi.org/10.1016/j.jclinepi.2018.02.025>

Reuse

This article is distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs (CC BY-NC-ND) licence. This licence only allows you to download this work and share it with others as long as you credit the authors, but you can't change the article in any way or use it commercially. More information and the full terms of the licence here: <https://creativecommons.org/licenses/>

Takedown

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

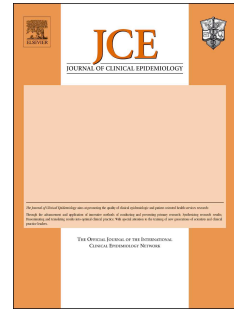


eprints@whiterose.ac.uk
<https://eprints.whiterose.ac.uk/>

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](https://doi.org/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.

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

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.

1 Background

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

10
11 In this systematic review, we seek to identify the metrics or methods used to calculate the
12 effectiveness of literature searching in health and allied topics. We also seek to explore if
13 the metrics or methods are used formatively or summatively (that is, do they seek to
14 predict or to evaluate effectiveness (see Figure 1). This study extends beyond simply
15 documenting how the effectiveness of literature searching has been calculated to
16 conducting a broader examination of what effectiveness means and how it might be
17 defined.

18 Methods

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

21
22 Research questions:

- 23 1: What metrics or methods are used to calculate literature search effectiveness?
- 24 2: Which metrics or methods are used formatively or summatively?
- 25 3: How is effectiveness defined in the studies?

26
27 Identifying studies and study data

28 29 *Searching bibliographic databases*

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

42 43 Study selection

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

46
47 The following inclusion criteria were applied hierarchically:

48
49 An original study published in the peer-reviewed literature that:

- 50
51 1. calculated literature search effectiveness;
52 2. provided sufficient information to replicate the calculation; and
53 3. calculated effectiveness between a supplementary search method (e.g.
54 handsearching, citation chasing, web searching, contacting study authors or trials
55 register searching) and bibliographic database searching.

56
57 The following studies were excluded:

- 58
59 • studies which did not compare the effectiveness of a supplementary search method
60 against bibliographic database searching;
61 • studies evaluating effectiveness of teaching literature searching (i.e. trained vs.
62 novice literature searchers);
63 • studies evaluating only search filters (i.e. 'search filter (a)' was compared to 'search
64 filter (b)');
65 • studies evaluating the effectiveness of tools (i.e. Google Scholar vs. Web of Science);
66 and
67 • abstracts, non-English language papers, letters, reviews and incomplete studies (i.e.
68 those which do not report effectiveness outcomes).

69
70 Data extraction

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

73
74 The following data were extracted: study citation, reference standard index test metric(s)
75 or method(s) to calculate effectiveness, definition of effectiveness reported in the study (i.e.
76 threshold), and claimed advantages and disadvantages relating to the calculation of
77 effectiveness. Data were also extracted if search strategies for a reference or index test
78 were reported and if methods to validate or quality appraise the reference standard or
79 index test were reported. Furthermore, we determined if the evaluation was derived
80 formatively (the purpose of the evaluation was to estimate) or summatively (the purpose of
81 the evaluation was to calculate). The following terms are defined in figure one: reference
82 standard, index test, summative and formative.

83
84 Quality assessment

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

88
89 Data synthesis

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

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

98 Results

99 Database searching identified 9,126 studies for title/abstract screening after de-duplication.
100 200 studies were screened at full-text and 50 studies met the inclusion criteria. The
101 Preferred Reporting in Systematic Reviews and Meta-Analysis (PRISMA) flow diagram is
102 recorded in figure 2 (2) and studies excluded at full-text are identified in supplementary
103 material.
104

105 Study characteristics

106 Of the 50 included studies (**Error! Reference source not found.**), 46 (92%) used
107 handsearching as the reference standard. The remaining four studies used another review
108 (n=1) or a specific combination of database searching (n=3). Validating the method or
109 searches used to develop the reference standard was reported in 26 of 50 studies (52%) and
110 to develop the index test in three of 50 studies (3%). Identifying a threshold to test
111 effectiveness against was reported in 17 of 50 studies (34%). Confidence intervals were
112 reported in 52% (26 of 50) of studies
113

114 Research Question 1 and 2: what metrics and methods are used to measure 115 literature search effectiveness and which metrics or methods are formative or 116 summative?

117 The metrics and methods used to calculate effectiveness (including specific equations) are
118 reported in figure 3. Six metrics and one method used to calculate and evaluate literature
119 search effectiveness were identified and had been used either individually or in
120 combination. These metrics and methods are summarised narratively below and the
121 calculations are reported in Table 1.
122

123 Six Metrics: summative

124

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

126 Sensitivity refers to the proportion of studies correctly identified as relevant, relative to the
127 total number of relevant studies that may exist. All 45 studies evaluating sensitivity used
128 the same calculation to determine a value, although the calculations are reported
129 differently according to the type of study in which they are used (figure 3). Sensitivity is also
130 referred to as: Recall (9, 21, 47) or relative recall¹.
131

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..

132 Specificity: 34/50 (68%) studies identified (4, 7, 8, 11-18, 23-26, 28-32, 34-47). Specificity
133 refers to the number of irrelevant studies excluded or not identified by the literature search
134 strategy. All 34 studies evaluating specificity used the same metric to determine a value
135 (figure 3).
136

137 Precision: 40/50 (80%) studies identified (3-5, 7, 9-17, 20-26, 28-47)
138 Precision refers to the number of relevant studies identified by a literature search. All 40
139 studies used the same metric to determine a value (figure 3). Precision was also referred to
140 as: Positive predictive value (or PPV (4, 17)).
141

142 Accuracy: 22/50 (44%) studies identified (11-16, 23-25, 30-32, 34-38, 40, 44-47) Accuracy
143 refers to the proportion of all studies correctly identified compared to the number of non-
144 relevant studies. All 22 studies used the same metric to determine a value (figure 3).
145

146 Number Needed to Read (NNR): 8/50 (16%) studies identified (5, 8, 9, 20, 28, 31, 41, 49).
147 NNR is defined as the number of studies a researcher has to read to identify a relevant
148 study. All 7 studies used the same metric to determine a value (figure 3). NNR was also
149 referred to as: Number Needed to Search (28).
150

151 Yield (summative): 4/50 (8%) studies identified (10, 50-52)
152 Yield refers to the number of studies identified by a literature search method. All 4 studies
153 interpreted yield in the same way.
154

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

158 **One Method: formative**
159

160 Capture-Recapture (Population Estimate): 2/50 (4%) studies identified (19, 53)
161 Capture-Recapture (or capture mark recapture) is a formative method which provides an
162 estimate of the 'population' of potentially relevant studies that might meet inclusion
163 criteria.
164

165 Combinations of the above methods were commonly used. These combinations are
166 summarised in Table 1.
167

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

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

176 Discussion

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

182

183 Formative methods

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

190

191 Summative methods

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

197

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

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

208

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

222

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).

270
271 Sampson et. al have proposed a method ('Inquisitio validus Index Medicus') for search
272 validation (54), and the Peer Review of Electronic Search Strategies (PRESS Checklist)
273 exists for the review of electronic search methods (76). A study by Hausner et al recorded
274 the time taken to quality appraise searches used in effectiveness evaluation as between 0.5
275 to 6.75 hours (77). Reporting the validation of methods used to develop reference standard
276 or index tests, and their corresponding searches, should be a particular focus of studies
277 seeking to calculate or estimate effectiveness of literature searching. Errors generated in
278 producing a 'test set' will necessarily impact on the accuracy of their effectiveness estimate.

279 **Use of thresholds**

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

289
290 Diagnostic tests determine and report thresholds to indicate the point at which results are
291 classified as either negative or positive (59). The prospective and clear reporting of
292 thresholds in evaluation studies of search strategies would aid interpretation of the studies
293 and would inform corresponding estimates of effectiveness generally, if the reporting of
294 thresholds was clearer. Glanville et al prospectively determined 'ideal performance' levels
295 for search filters through discussion with the project team. Whilst these levels were not
296 realised within the study, their evaluation of literature search effectiveness was
297 consequently easier to understand and analyse relative to their objectives (78).

298 **Confidence intervals**

299
300 Confidence intervals were reported in 52% of studies. Confidence intervals offer the reader
301 an estimate of certainty (and conversely of uncertainty) in connection with the estimate of
302 effect. Confidence intervals should, in our opinion, be calculated and reported in all studies
303 that seek to calculate search effectiveness.

304 **Sample size**

305
306 Harbour et al reported that sample size calculations were not reported in their evaluation of
307 search filter performance and our study shares similar conclusions (1). The number of
308 studies included in the reference standard impacts upon the reliability of the effectiveness
309 estimate. The reporting of sample size calculations, or alternatively why it was not
310 considered possible to generate a reliable sample, is recommended.

311 **Value**

312
313 Effectiveness, reported in purely quantitative terms, tells researchers little about the value
314 of the studies identified or missed, or what the effect of missing studies means (60). It is
315 unclear what proportion of relevant studies identified represents an adequate literature
316 search, so researchers are presently required to make their own judgements of sensitivity

317 (79-81). Sensitivity values do not help researchers understand this problem. It is
318 acknowledged that no search can record 100% sensitivity (82, 83), so what does a 90%
319 value demonstrate, other than that 10% of studies might be missing? Determining steps to
320 identify the missing 10% (where comprehensive study identification is important to the
321 review), or why a search was stopped, would be of benefit when reporting literature
322 searches (84). The more pressing issue appears to be whether to revisit assumptions of the
323 usefulness of evaluating literature searches by measuring comprehensiveness, since
324 comprehensiveness may not be an appropriate indicator of search quality (82).

325

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

337

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

346 Limitations

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

353

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

358 Conclusions

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

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

363
364 Studies evaluating effectiveness need to identify clearly the threshold at which they will
365 define effectiveness and how the evaluation they report correlates to this threshold. We
366 found that this is not yet common practice.

367
368 Studies that attempt to investigate literature search effectiveness should be informed by
369 the reporting of confidence intervals, which aids interpretation of uncertainty within the
370 result, and the search methods used to derive effectiveness estimates should be clearly
371 reported and clearly validated in studies.

ACCEPTED MANUSCRIPT

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

374
375 **Funding:** This work forms a chapter of Chris Cooper's PhD. Chris' PhD was funded by an
376 NIHR Health Technology Grant held at the University of Exeter.

377
378 RG and NB were partially supported by the National Institute for Health Research (NIHR)
379 Collaboration for Leadership in Applied Health Research and Care South West Peninsula.

380
381 The views expressed are those of the author(s) and not necessarily those of the NHS, the
382 NIHR or the Department of Health.

383
384 References

- 385
386 1. Harbour J, Fraser C, Lefebvre C, Glanville J, Beale S, Boachie C, et al. Reporting
387 methodological search filter performance comparisons: a literature review. *Health*
388 *Information & Libraries Journal*. 2014;31(3):176-94.
- 389 2. Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic
390 reviews and meta-analyses: The PRISMA statement. *Journal of Clinical Epidemiology*.
391 2009;62:1006-12.
- 392 3. Adams CE, Power A, Frederick K, Lefebvre C. An investigation of the adequacy of
393 MEDLINE searches for randomized controlled trials (RCTs) of the effects of mental health
394 care. *Psychological Medicine*. 1994;24(3):741-8.
- 395 4. Astin MP, Brazzelli MG, Fraser CM, Counsell CE, Needham G, Grimshaw JM.
396 Developing a sensitive search strategy in MEDLINE to retrieve studies on assessment of the
397 diagnostic performance of imaging techniques. *Radiology*. 2008;247(2):365-73.
- 398 5. Bachmann LM, Coray R, Estermann P, Ter Riet G. Identifying diagnostic studies in
399 MEDLINE: reducing the number needed to read. *Journal of the American Medical*
400 *Informatics Association*. 2002;9(6):653-8.
- 401 6. Cathey J, Al Hajeri AA, Fedorowicz Z. A comparison of handsearching versus
402 EMBASE searching of the Annals of Saudi Medicine to identify reports of randomized
403 controlled trials. *Annals of Saudi Medicine*. 2006;26(1):49-51.
- 404 7. Dumbrigue HB, Esquivel JF, Jones JS. Assessment of MEDLINE search strategies
405 for randomized controlled trials in prosthodontics. *Journal of Prosthodontics*. 2000;9(1):8-13.
- 406 8. Geersing GJ, Bouwmeester W, Zuithoff P, Spijker R, Leeflang M, Moons K. Search
407 filters for finding prognostic and diagnostic prediction studies in medline to enhance
408 systematic reviews. *PLoS ONE*. 2012;7 (2) (no pagination)(e32844).
- 409 9. Gehanno JF, Rollin L, Le Jean T, Louvel A, Darmoni S, Shaw W. Precision and recall
410 of search strategies for identifying studies on return-to-work in Medline. *Journal of*
411 *Occupational Rehabilitation*. 2009;19(3):223-30.
- 412 10. Glanville JM, Duffy S, McCool R, Varley D. Searching ClinicalTrials.gov and the
413 International Clinical Trials Registry Platform to inform systematic reviews: what are the
414 optimal search approaches? *Journal of the Medical Library Association*. 2014;102(3):177-83.
- 415 11. Haynes RB, Kastner M, Wilczynski NL. Developing optimal search strategies for
416 detecting clinically sound and relevant causation studies in EMBASE. *BMC Medical*
417 *Informatics and Decision Making*. 2005;5 (no pagination)(8).
- 418 12. Haynes RB, McKibbon KA, Wilczynski NL, Walter SD, Werre SR. Optimal search
419 strategies for retrieving scientifically strong studies of treatment from Medline: Analytical
420 survey. *British Medical Journal*. 2005;330(7501):1179-82.

- 421 13. Haynes RB, Wilczynski N, McKibbin KA, Walker CJ, Sinclair JC. Developing
422 optimal search strategies for detecting clinically sound studies in MEDLINE. *Journal of the*
423 *American Medical Informatics Association*. 1994;1(6):447-58.
- 424 14. Haynes RB, Wilczynski NL. Optimal search strategies for retrieving scientifically
425 strong studies of diagnosis from Medline: analytical survey. *BMJ*. 2004;328(7447):1040.
- 426 15. Hildebrand AM, Iansavichus AV, Haynes RB, Wilczynski NL, Mehta RL, Parikh CR,
427 et al. High-performance information search filters for acute kidney injury content in PubMed,
428 Ovid Medline and Embase. *Nephrology Dialysis Transplantation*. 2014;29(4):823-32.
- 429 16. Holland JL, Wilczynski NL, Haynes RB. Optimal search strategies for identifying
430 sound clinical prediction studies in EMBASE. *BMC medical informatics and decision*
431 *making*. 2005;5:11.
- 432 17. Ingui BJ, Rogers MA. Searching for clinical prediction rules in MEDLINE. *Journal of*
433 *the American Medical Informatics Association : JAMIA*. 2001;8(4):391-7.
- 434 18. Jenuwine ES, Floyd JA. Comparison of Medical Subject Headings and text-word
435 searches in MEDLINE to retrieve studies on sleep in healthy individuals. *Journal of the*
436 *Medical Library Association*. 2004;92(3):349-53.
- 437 19. Kassaï B, Sonié S, Shah NR, Boissel J-P. Literature search parameters marginally
438 improved the pooled estimate accuracy for ultrasound in detecting deep venous thrombosis.
439 *Journal of clinical epidemiology*. 2006;59(7):710-4.
- 440 20. Layton DM, Clarke M. Search Strategy to Identify Dental Survival Analysis Articles
441 Indexed in MEDLINE. *International Journal of Prosthodontics*. 2016;29(1):20-7.
- 442 21. Linder SK, Kamath GR, Pratt GF, Saraykar SS, Volk RJ. Citation searches are more
443 sensitive than keyword searches to identify studies using specific measurement instruments.
444 *Journal of Clinical Epidemiology*. 2015;68(4):412-7.
- 445 22. Marson AG, Chadwick DW. How easy are randomized controlled trials in epilepsy to
446 find on Medline? The sensitivity and precision of two Medline searches. *Epilepsia*.
447 1996;37(4):377-80.
- 448 23. McKibbin KA, Wilczynski NL, Haynes RB. Developing optimal search strategies for
449 retrieving qualitative studies in PsycINFO. *Evaluation & the Health Professions*.
450 2006;29(4):440-54.
- 451 24. McKibbin KA, Wilczynski NL, Haynes RB. Retrieving randomized controlled trials
452 from medline: A comparison of 38 published search filters. *Health Information and Libraries*
453 *Journal*. 2009;26(3):187-202.
- 454 25. McKinlay RJ, Wilczynski NL, Haynes RB. Optimal search strategies for detecting
455 cost and economic studies in EMBASE. *BMC Health Services Research*. 2006;6 (no
456 pagination)(67).
- 457 26. Montori VM, Wilczynski NL, Morgan D, Haynes RB. Optimal search strategies for
458 retrieving systematic reviews from Medline: Analytical survey. *British Medical Journal*.
459 2005;330(7482):68-71.
- 460 27. Nasser M, Al Hajeri A. A comparison of handsearching versus embase searching of
461 the archives of Iranian medicine to identify reports of randomized controlled trials. *Archives*
462 *of Iranian Medicine*. 2006;9(3):192-5.
- 463 28. Rogerson TE, Ladhani M, Mitchell R, Craig JC, Webster AC. Efficient strategies to
464 find diagnostic test accuracy studies in kidney journals. *Nephrology*. 2015;20(8):513-8.
- 465 29. Taljaard M, McGowan J, Grimshaw JM, Brehaut JC, McRae A, Eccles MP, et al.
466 Electronic search strategies to identify reports of cluster randomized trials in MEDLINE: low
467 precision will improve with adherence to reporting standards. *BMC Medical Research*
468 *Methodology*. 2010;10:15.

- 469 30. Ugolini D, Neri M, Casilli C, Bonassi S. Development of search filters for retrieval of
470 literature on the molecular epidemiology of cancer. *Mutation Research - Genetic Toxicology*
471 *and Environmental Mutagenesis*. 2010;701(2):107-10.
- 472 31. van de Glind EM, van Munster BC, Spijker R, Scholten RJ, Hooft L. Search filters to
473 identify geriatric medicine in Medline. *Journal of the American Medical Informatics*
474 *Association*. 2012;19(3):468-72.
- 475 32. Walters LA, Wilczynski NL, Haynes RB. Developing optimal search strategies for
476 retrieving clinically relevant qualitative studies in EMBASE. *Qualitative Health Research*.
477 2006;16(1):162-8.
- 478 33. Watson RJ, Richardson PH. Identifying randomized controlled trials of cognitive
479 therapy for depression: comparing the efficiency of Embase, Medline and PsycINFO
480 bibliographic databases. *The British journal of medical psychology*. 1999;72 (Pt 4):535-42.
- 481 34. Wilczynski NL, Haynes RB. Developing optimal search strategies for detecting
482 clinically sound causation studies in MEDLINE. *Amia 2003;Annual Symposium proceedings*
483 */ AMIA Symposium*. AMIA Symposium.:719-23.
- 484 35. Wilczynski NL, Haynes RB. Optimal search strategies for detecting clinically sound
485 prognostic studies in EMBASE: an analytic survey. *Journal of the American Medical*
486 *Informatics Association*. 2005;12(4):481-5.
- 487 36. Wilczynski NL, Haynes RB. EMBASE search strategies achieved high sensitivity and
488 specificity for retrieving methodologically sound systematic reviews. *Journal of Clinical*
489 *Epidemiology*. 2007;60(1):29-33.
- 490 37. Wilczynski NL, Haynes RB, Eady A, Haynes B, Marks S, McKibbin A, et al.
491 Developing optimal search strategies for detecting clinically sound prognostic studies in
492 MEDLINE: An analytic survey. *BMC Medicine*. 2004;2 (no pagination)(23).
- 493 38. Wilczynski NL, Haynes RB, Hedges T. Optimal search strategies for identifying
494 mental health content in MEDLINE: An analytic survey. *Annals of General Psychiatry*.
495 2006;5 (no pagination)(4).
- 496 39. Wilczynski NL, Haynes RB, Lavis JN, Ramkissoonsingh R, Arnold-Oatley AE.
497 Optimal search strategies for detecting health services research studies in MEDLINE. *Cmaj*.
498 2004;171(10):1179-85.
- 499 40. Wilczynski NL, Haynes RB, Team QIH. Optimal search filters for detecting quality
500 improvement studies in Medline. *Quality & Safety in Health Care*. 2010;19(6):e31.
- 501 41. Wilczynski NL, McKibbin KA, Haynes RB. Search filter precision can be improved
502 by NOTing out irrelevant content. *AMIA Annual Symposium Proceedings/AMIA*
503 *Symposium*. 2011;2011:1506-13.
- 504 42. Wilczynski NL, Walker CJ, McKibbin KA, Haynes RB. Assessment of methodologic
505 search filters in MEDLINE. *Proceedings - the Annual Symposium on Computer Applications*
506 *in Medical Care*. 1993:601-5.
- 507 43. Wilczynski NL, Walker CJ, McKibbin KA, Haynes RB. Quantitative comparison of
508 pre-exposures and subheadings with methodologic search terms in MEDLINE. *Proceedings*
509 *Symposium on Computer Applications in Medical Care*. 1994:905-9.
- 510 44. Wong SS, Wilczynski NL, Haynes RB. Developing optimal search strategies for
511 detecting clinically relevant qualitative studies in MEDLINE. *Medinfo*. 2004;MEDINFO.
512 11(Pt 1):311-6.
- 513 45. Wong SS, Wilczynski NL, Haynes RB. Developing optimal search strategies for
514 detecting clinically sound treatment studies in EMBASE. *Journal of the Medical Library*
515 *Association*. 2006;94(1):41-7.
- 516 46. Wong SS, Wilczynski NL, Haynes RB. Optimal CINAHL search strategies for
517 identifying therapy studies and review articles. *Journal of Nursing Scholarship*.
518 2006;38(2):194-9.

- 519 47. Wong SS, Wilczynski NL, Haynes RB, Ramkissoonsingh R. Developing optimal
520 search strategies for detecting sound clinical prediction studies in MEDLINE. *Amia*
521 2003;Annual Symposium proceedings / AMIA Symposium. AMIA Symposium.:728-32.
- 522 48. Eysenbach G, Tuische J, Diepgen TL. Evaluation of the usefulness of Internet
523 searches to identify unpublished clinical trials for systematic reviews. *Medical informatics*
524 *and the Internet in medicine*. 2001;26(3):203-18.
- 525 49. Mattioli S, Farioli A, Cooke RM, Baldasseroni A, Ruotsalainen J, Placidi D, et al.
526 Hidden effectiveness? Results of hand-searching Italian language journals for occupational
527 health interventions. *Occupational & Environmental Medicine*. 2012;69(7):522-4.
- 528 50. Blanc X, Collet TH, Auer R, Iriarte P, Krause J, Legare F, et al. Retrieval of
529 publications addressing shared decision making: an evaluation of full-text searches on
530 medical journal websites. *JMIR Research Protocols*. 2015;4(2):e38.
- 531 51. Glanville J, Cikaló M, Crawford F, Dozier M, McIntosh H. Handsearching did not
532 yield additional unique FDG-PET diagnostic test accuracy studies compared with electronic
533 searches: a preliminary investigation. *Research Synthesis Methods*. 2012;3(3):202-13.
- 534 52. Hopewell S, Clarke M, Lusher A, Lefebvre C, Westby M. A comparison of
535 handsearching versus MEDLINE searching to identify reports of randomized controlled trials.
536 *Statistics in Medicine*. 2002;21(11):1625-34.
- 537 53. Spoor P, Airey M, Bennett C, Greensill J, Williams R. Use of the capture-recapture
538 technique to evaluate the completeness of systematic literature searches. *BMJ*.
539 1996;313(7053):342-3.
- 540 54. Sampson M, McGowan J. Inquisitio validus Index Medicus: A simple method of
541 validating MEDLINE systematic review searches. *Research Synthesis Methods*.
542 2011;2(2):103-9.
- 543 55. Kok R, Verbeek JAHM, Faber B, van Dijk FJH, Hoving JL. A search strategy to
544 identify studies on the prognosis of work disability: a diagnostic test framework. *BMJ Open*.
545 2015;5(5).
- 546 56. Waffenschmidt S, Guddat C. Searches for randomized controlled trials of drugs in
547 MEDLINE and EMBASE using only generic drug names compared with searches applied in
548 current practice in systematic reviews. *Res Synth Methods*. 2015;6(2):188-94.
- 549 57. Lee E, Dobbins M, Decorby K, McRae L, Tirilis D, Husson H. An optimal search
550 filter for retrieving systematic reviews and meta-analyses. *BMC Med Res Methodol*.
551 2012;12:51.
- 552 58. Jenkins M. Evaluation of methodological search filters--a review. *Health Info Libr J*.
553 2004;21(3):148-63.
- 554 59. Centre for R, Dissemination. Systematic reviews – CRD’s guidance for undertaking
555 reviews in healthcare. [Edition no.] ed. York: Centre for Reviews and Dissemination,
556 University of York; 2009.
- 557 60. Cooper C, Booth A, Britten N, Garside R. A comparison of results of empirical
558 studies of supplementary search techniques and recommendations in review methodology
559 handbooks: a methodological review. *Systematic Reviews*. 2017;6(1):234.
- 560 61. Adams CE, Power A, Frederick K, Lefebvre C. An investigation of the adequacy of
561 MEDLINE searches for randomized controlled trials (RCTs) of the effects of mental health
562 care. *Psychol Med*. 1994;24(3):741-8.
- 563 62. Armstrong R, Jackson N, Doyle J, Waters E, Howes F. It’s in your hands: the value of
564 handsearching in conducting systematic reviews of public health interventions. *Journal of*
565 *Public Health*. 2005;27(4):388-91.
- 566 63. Croft AM, Vassallo DJ, Rowe M. Handsearching the Journal of the Royal Army
567 Medical Corps for trials. *Journal of the Royal Army Medical Corps*. 1999;145(2):86-8.

- 568 64. Hay PJ, Adams CE, Lefebvre C. The efficiency of searches for randomized controlled
569 trials in the International Journal of Eating Disorders: a comparison of handsearching,
570 EMBASE and PsycLIT. *Health Libraries Review*. 1996;13(2):91-6.
- 571 65. Hopewell S, Clarke M, Lusher A, Lefebvre C, Westby M. A comparison of
572 handsearching versus MEDLINE searching to identify reports of randomized controlled trials.
573 *Stat Med*. 2002;21(11):1625-34.
- 574 66. Jadad AR, McQuay HJ. A high-yield strategy to identify randomized controlled trials
575 for systematic reviews. *The Online journal of current clinical trials*. 1993;Doc No 33:[3973
576 words; 39 paragraphs].
- 577 67. Langham J, Thompson E, Rowan K. Identification of randomized controlled trials
578 from the emergency medicine literature: comparison of hand searching versus MEDLINE
579 searching. *Ann Emerg Med*. 1999;34(1):25-34.
- 580 68. Sampson M, Zhang L, Morrison A, Barrowman NJ, Clifford TJ, Platt RW, et al. An
581 alternative to the hand searching gold standard: validating methodological search filters using
582 relative recall. *BMC Med Res Methodol*. 2006;6:33.
- 583 69. Doust JA, Pietrzak E, Sanders S, Glasziou PP. Identifying studies for systematic
584 reviews of diagnostic tests was difficult due to the poor sensitivity and precision of
585 methodologic filters and the lack of information in the abstract. *Journal of Clinical
586 Epidemiology*. 2005;58(5):444-9.
- 587 70. Ugolini D, Neri M, Casilli C, Bonassi S. Development of search filters for retrieval of
588 literature on the molecular epidemiology of cancer. *Mutation research*. 2010;701(2):107-10.
- 589 71. Vlachos A, Craven M. Biomedical event extraction from abstracts and full papers
590 using search-based structured prediction. *BMC Bioinformatics*. 2012;13(Suppl 11):S5-S.
- 591 72. Guns R, Lioma C, Larsen B. The tipping point: F-score as a function of the number of
592 retrieved items. *Information Processing & Management*. 2012;48(6):1171-80.
- 593 73. Hripcsak G, Rothschild AS. Agreement, the F-Measure, and Reliability in
594 Information Retrieval. *Journal of the American Medical Informatics Association*.
595 2005;12(3):296-8.
- 596 74. Liu Z, Tan M, Jiang F. Regularized F-measure maximization for feature selection and
597 classification. *Journal of Biomedicine & Biotechnology*. 2009;2009:617946.
- 598 75. Patrick TB, Demiris G, Folk LC, Moxley DE, Mitchell JA, Tao D. Evidence-based
599 retrieval in evidence-based medicine. *Journal of the Medical Library Association*.
600 2004;92(2):196-9.
- 601 76. McGowan J, Sampson M, Salzwedel DM, Cogo E, Foerster V, Lefebvre C. PRESS
602 Peer Review of Electronic Search Strategies: 2015 Guideline Statement. *Journal of Clinical
603 Epidemiology*. 2016;75:40-6.
- 604 77. Hausner E, Guddat C, Hermanns T, Lampert U, Waffenschmidt S. Development of
605 search strategies for systematic reviews: validation showed the noninferiority of the objective
606 approach. *J Clin Epidemiol*. 2015;68(2):191-9.
- 607 78. Glanville J, Kaunelis D, Mensinkai S. How well do search filters perform in
608 identifying economic evaluations in MEDLINE and EMBASE. *Int J Technol Assess Health
609 Care*. 2009;25(4):522-9.
- 610 79. Lane D, Dykeman J, Ferri M, Goldsmith CH, Stelfox HT. Capture-mark-recapture as
611 a tool for estimating the number of articles available for systematic reviews in critical care
612 medicine. *Journal of Critical Care*. 2013;28(4):469-75.
- 613 80. Kastner M, Straus SE, McKibbin KA, Goldsmith CH. The capture-mark-recapture
614 technique can be used as a stopping rule when searching in systematic reviews. *J Clin
615 Epidemiol*. 2009;62(2):149-57.
- 616 81. Siva S. Optimal strategies for literature search. *Indian Journal of Urology : IJU :
617 Journal of the Urological Society of India*. 2009;25(2):246-50.

- 618 82. Booth A. How much searching is enough? Comprehensive versus optimal retrieval
619 for technology assessments. *Int J Technol Assess Health Care*. 2010;26(4):431-5.
- 620 83. Hildebrand AM, Iansavichus AV, Lee CW, Haynes RB, Wilczynski NL, McKibbin
621 KA, et al. Glomerular disease search filters for Pubmed, Ovid Medline, and Embase: a
622 development and validation study. *BMC Med Inform Decis Mak*. 2012;12:49.
- 623 84. Webster AJ, Kemp R. Estimating Omissions From Searches. *The American*
624 *Statistician*. 2013;67(2):82-9.
- 625 85. Ross-White A, Godfrey C. Is there an optimum number needed to retrieve to justify
626 inclusion of a database in a systematic review search? *Health Info Libr J*. 2017.
- 627 86. Egan M, MacLean A, Sweeting H, Hunt K. Comparing the effectiveness of using
628 generic and specific search terms in electronic databases to identify health outcomes for a
629 systematic review: a prospective comparative study of literature search methods. *BMJ Open*.
630 2012;2(3).
- 631 87. Booth A. The number needed to retrieve: a practically useful measure of information
632 retrieval? *Health Info Libr J*. 2006;23(3):229-32.
- 633

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	Identify 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

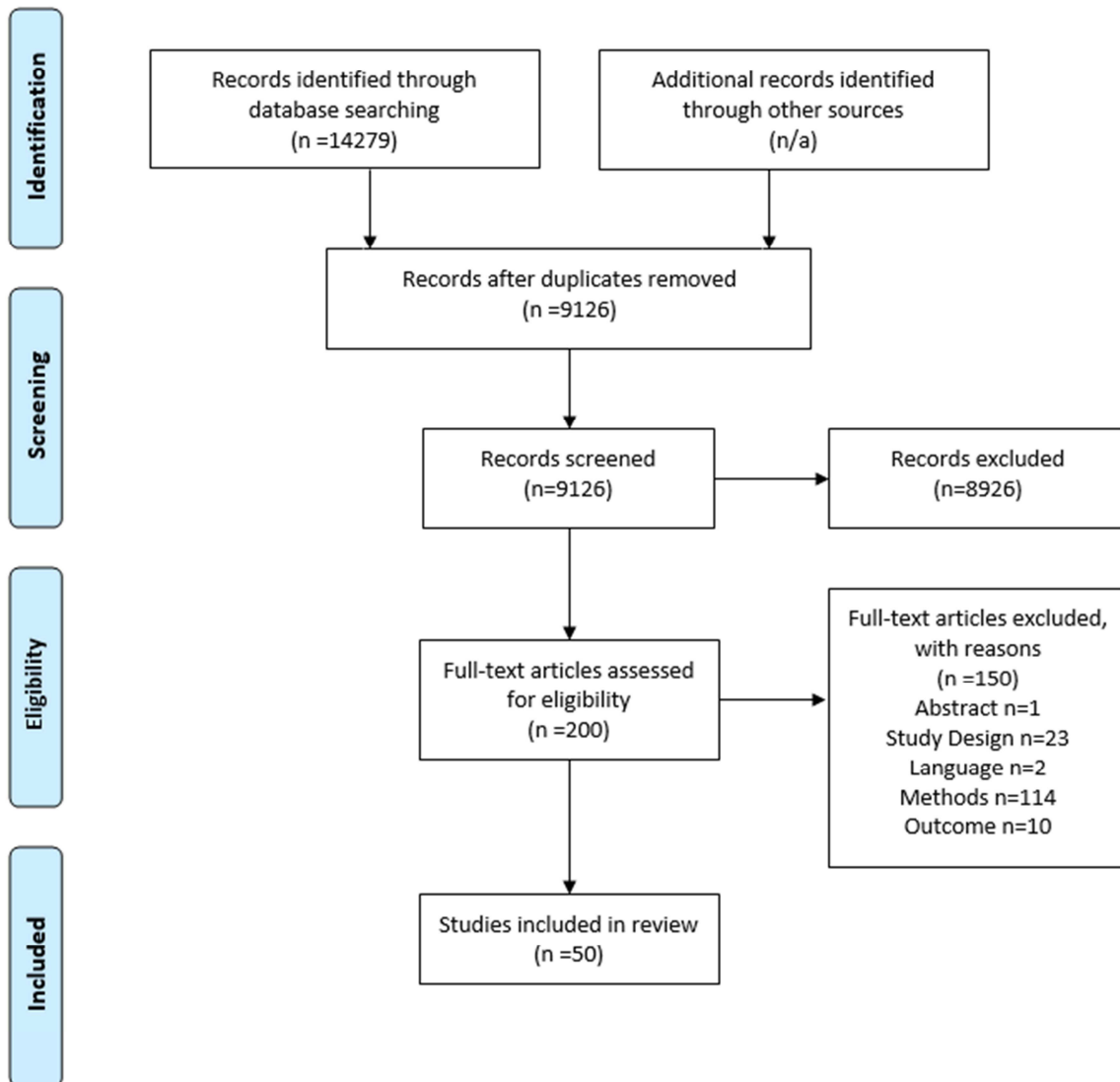


Figure 1 schematic of key metrics and methods to evaluate literature search effectiveness and their respective calculations

Index Test	Reference Standard			
	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	$\frac{\text{number of article by search method A}}{\text{number of articles by search method A+B}} \times \frac{\text{number of articles by search method B}}{\text{number of articles by search method A+B}}$		$x (y/z)$

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'

ACCEPTED MANUSCRIPT