Reporting methodological search filter performance comparisons: a literature review

Authors

Jennifer Harbour, Healthcare Improvement Scotland, Delta House, 50 West Nile Street, Glasgow G1 2NP. Email: jenny.harbour@nhs.net

Cynthia Fraser, Health Services Research Unit, Institute of Applied Health Sciences, University of Aberdeen.

Carol Lefebvre, Lefebvre Associates Ltd, Oxford.

Julie Glanville, York Health Economics Consortium Ltd, York.

Sophie Beale, University of Liverpool, Liverpool.

Charles Boachie, Health Services Research Unit, Institute of Applied Health Sciences, University of Aberdeen.

Steven Duffy, Centre for Reviews and Dissemination, University of York, York.

Rachael McCool, York Health Economics Consortium Ltd, York.

Lynne Smith, Healthcare Improvement Scotland, Glasgow.

Acknowledgements

The authors would like to acknowledge the funding for this review as part of a wider project to explore the methods used in selecting and assessing the performance of search filters within evidence based health research. Funding was obtained from the UK Medical Research Council under the MRC-NIHR Methodology Research Programme to support NICE decision-making (grant number G0901496).

The authors would also like to acknowledge contributions from Danielle Varley (York Health Economics Consortium, York) who has supported and commented on the production of the article and other publications generated from the wider MRC-funded project.

Key Messages

- When selecting a filter consider whether you need a filter with high sensitivity or high precision since these are the measures most frequently reported.
- Studies that compare search filter performance should explicitly report methods and results to help searchers identify the most appropriate filter.
- “Translating” search filters between databases or database interfaces should be done carefully and the changes made should be recorded accurately.
- Studies presenting the development of new search filters that include comparisons with existing filters should present detailed methods describing how the performance comparisons were undertaken.
- One or more clearly described gold standards should be used to test comparative filter performance.

Keywords: Databases, bibliographic; Information storage and retrieval; Review, literature; Sensitivity and specificity.

Abstract

Background

Methodological search filters are tools for retrieving database records reporting studies which use a specific research method. Choosing a filter is likely to be based on filter performance data. This review examines which measures are reported, and the way that filter performance is presented, in filter comparisons.

Methods

Studies were identified from the current content and pending update (2010) of a filter website. Eligible studies compared two or more methodological search filters designed to identify randomized controlled trials, diagnostic test accuracy studies, systematic reviews or economic evaluations.

Results

Eighteen studies met the inclusion criteria. The number of filters compared in a single study ranged from 2 to 38. The most commonly reported measures were sensitivity/recall and precision. All studies displayed results in tables and gave results as percentages or proportions. Two studies supplemented results tables with graphical displays of data: a bar graph of the proportion of retrieved and missed gold standard references per filter; a forest plot of the overall sensitivity and specificity of each filter.

Conclusions

Sensitivity/recall and precision are the most frequently reported performance measures. This review highlights the potential for presenting results in novel and innovative ways to aid filter selection.

Background

The effective retrieval of published and unpublished literature is essential for developing clinical guidance, conducting health research, developing health policy and supporting healthcare decision making. The aim of evidence retrieval is to provide appropriate volumes of relevant information within the time and cost restraints that exist. Effective evidence retrieval should provide a robust set of results that can be used to establish accurate estimates of parameters such as clinical effectiveness and cost effectiveness, and minimize any bias that might be introduced through incomplete retrieval. Whether the purpose of the evidence retrieval is to find a representative set of results to inform the development of an economic model or to conduct an extensive search for evidence on the effects of a healthcare intervention, retrieval methods need to be appropriate, efficient, consistent and reliable.

One tool which is widely used by information professionals, researchers and others engaged in finding clinical evidence is the search filter. Search filters seek to capture a search concept. The search concept may be a study design, such as randomized controlled trials, an aspect of research such as adverse events, a population such as children, or a disease/condition such as Parkinson's disease. A methodological search filter is a combination of search terms designed to identify records of studies that have used a specific research method. Effective search filters may seek to optimize retrieval using a balance between maximizing sensitivity (identifying as high a proportion as possible of relevant records) and achieving adequate precision (minimizing the number of irrelevant records), or they may seek to maximize sensitivity or precision only. Using well-designed, relevant search filters should offer a standard approach to study retrieval and release searcher time to focus on developing other aspects of the information retrieval task, such as the most appropriate terms for identifying studies on a specific illness, of a particular treatment or with certain patient outcomes.

A variety of methodological search filters are already available to find randomized controlled trials, economic evaluations, systematic reviews and many other study designs. In principle, these filters can offer efficient, validated and consistent approaches to study identification within large bibliographic databases. However, search filters are an under-researched tool. Although there are many published search filters, few are extensively validated beyond the data offered in their original publication.[1-4] This means that their performance in the real-world setting of day-to-day information retrieval across a range of search topics is unknown.[5] Furthermore, search filters are seldom assessed against common datasets which makes comparison of performance across filters problematic. Consequently the use of search filters as a standard tool within technology assessment, guideline development and other evidence syntheses may be pragmatic rather than evidence-based.[5, 6]

As search filters proliferate, the key question becomes how to choose between them. The most useful information to assist search filter choice is likely to be performance data derived from well-conducted and well-reported performance tests or comparisons. Methods exist to test search filter performance and to build the performance picture,

including reviews of search filter performance.[1, 2, 7-9] However, there is no formal guidance on the best methods for testing filter performance, on which performance measures are valued by searchers and which measures should ideally be reported to assist searchers in choosing between filters. The performance picture for filters across different disciplines, questions and databases is therefore largely unknown. Different performance measures are reported in studies describing search filters, and the process whereby searchers choose a filter remains unclear.

In 2010 the Medical Research Council (MRC), in partnership with the National Institute for Health and Care Excellence (NICE), funded a research project (MRC research grant G0901496) to improve our understanding of search filter use, how searchers and researchers choose search filters, and what information searchers and researchers would like to receive to inform their choice of filter. This research involved a multi-method approach:

- Five literature reviews investigating different aspects of performance measurement in search filters and diagnostic test accuracy studies (to which they are analogous), their reporting and the selection of search filters by searchers and researchers;
- Interviews and a web-based questionnaire to gain information on current filter use;
- Development of examples of filter performance visualization and guidance on gathering and reporting search filter performance based on the reviews, interviews and questionnaire.

The purpose of the review reported in this article is to consider the measures and methods used in reporting the comparative performance of multiple methodological search filters as part of the above project.

Objectives

This review addresses the following questions:

- What performance measures are reported in studies comparing the performance of one or more methodological search filters in one or more sets of records?
- How are results presented in studies comparing the performance of one or more methodological search filters in one or more sets of records?
- How reliable are the methods used in studies comparing the performance of methodological search filters?
- Are there any published methods for synthesizing the results of several filter performance studies?
- Are there any published methods for reviewing the results of several syntheses?

Methods

Identification of studies

Potentially relevant studies were identified from the InterTASC Information Specialists' SubGroup (ISSG) Search Filter Resource (https://sites.google.com/a/york.ac.uk/issg-search-filters-resource/home) in 2010. The Search Filter Resource is a collaboratively produced, regularly updated, web resource listing published and unpublished search filters. Studies comparing the performance of one or more methodological search filters are also included in the Search Filter Resource.

Additional studies were identified from the results of an update search carried out in 2010 by the UK Cochrane Centre to support the ISSG Search Filter Resource. Studies which developed one or more filters and compared their performance to previously published filters were selected from the ISSG Search Filter Resource for a concurrent review and incorporated into this review (L. Smith, personal communication, September 2010).

Inclusion criteria

For the purpose of this review, methodological search filters were defined as '*any search filter or strategy used to identify database records of studies that use a particular clinical research method*'. Only studies comparing the performance of filters for randomized controlled trials (RCTs), diagnostic test accuracy studies, systematic reviews or economic evaluation studies were included.

Studies were selected for inclusion in the review if they compared the performance of two or more methodological search filters in one or more sets of records. Studies reporting the development of new methodological filters whose performance was compared with that of previously published filters were also included.

Exclusion criteria

Studies were excluded from the review if they:

- Reported the development and initial testing of a single search filter that did not include any formal comparison with the performance of other search filters.
- Compared methodological search filters that had not been designed to retrieve RCTs, diagnostic test accuracy studies, systematic reviews or economic evaluation studies.
- Compared the performance of a single filter in multiple databases or interfaces.
- Were not available as a full report, for example, conference abstracts.
- Were protocols for studies or reviews.
- Lacked sufficient methodological detail for the data extraction process.

Data extraction and synthesis

A data extraction form was developed by two reviewers (JH, CF) to standardize the extraction of data from the selected studies and allow cross-comparisons between studies. Details extracted included: the methods used to identify published filters for comparison; the methods used to test filter performance; and the performance measures reported. Data extraction for each study was carried out by one reviewer (JH) and verified by a second reviewer (CF). A narrative synthesis was used to summarize the results from the review.

Results

Twenty-one studies were identified as potentially meeting the inclusion criteria for this review based on the titles and abstracts.[1, 2, 10-28] Of these studies, ten reported the development of one or more search filters which were then compared against the performance of existing filters[19-28], and eleven reported comparative performance of existing filters.[1, 2, 10-18] On receipt of the full papers, three studies[10, 11, 14] were excluded from the review based on the criteria outlined in the methods section (Supplementary appendix 1). No studies were identified that synthesized the results of several performance reports or reviewed the results of several syntheses.

Characteristics of included studies

Of the eighteen studies included in the review:

- 8 reported the performance of diagnostic test accuracy search filters[1, 2, 12, 18, 20, 21, 27, 28]
- 5 reported the performance of RCT filters[13, 16, 19, 22, 23]
- 3 reported the performance of systematic review filters[24-26]
- 1 reported the performance of filters for economic evaluations[17]
- 1 reported the performance of RCT and systematic review filters.[15]

The methodological filters evaluated in the included studies had been developed in a variety of interfaces including LILACS, PubMed, Ovid and SilverPlatter. However, several studies did not specify the interface used in the development of some or all of the filters being compared.[2, 12, 13, 17-19, 21, 23-27] This absence of detail was particularly common in studies where performance comparison was secondary to the development of one or more new filters.[19, 21, 23-27]

Fourteen studies compared the performance of filters in MEDLINE (various platforms).[1, 2, 12, 13, 16, 18-20, 23-28] Two studies tested filters in MEDLINE and Embase.[15, 17] One study only tested Embase filters[21], and one study compared filters in LILACS.[22] Seven of the eight studies comparing diagnostic test accuracy filters used MEDLINE to test performance although the platform used varied.[1, 2, 12, 18, 20, 27, 28]

Studies included in the review used a variety of methods to identify relevant filters for comparison, including five which used database searches[1, 2, 13, 18, 20], four that consulted relevant websites[13, 17, 19, 20] and three that contacted experts in the field[2, 17, 18]. Ten

studies used other methods of identifying filters such as using studies they already knew about or studies they had conducted themselves.[2, 12, 13, 15, 18, 19, 23, 24, 27, 28] Five studies did not provide explicit details on how the filters for testing were identified.[16, 21, 22, 25, 26]

The number of filters compared in a single study ranged from 2 to 38. Diagnostic test accuracy study and RCT filters were the most common filters compared, and systematic review and economic evaluation filters the least common. Key characteristics of all included studies are summarized in Table 1 and further details are available in Supplementary appendix 2.

Table 1: Characteristics of performance comparison studies included in this review (full details in Supplementary appendix 2)

Study	How the filters identified for comparison?	What study type was the filter designed to retrieve?	Total number of included filters [number of included filters developed by the author]
Bachman (2002)[28]	Published filters	Diagnostic test accuracy studies	2 [1]
Boynton (1998)[26]	Published filters	Systematic reviews	15 [11]
Corrao (2006)[16]	Published filters, author modified strategy	RCTs	2
Deville (2000)[27]	Published filters	Diagnostic test accuracy studies	5 [4]
Doust (2005)[12]	Published filters	Diagnostic test accuracy studies	5
Glanville (2006)[23]	Published filters	RCTs	12 [6]
Glanville (2009)[17]	Websites, contacted experts	Economic evaluations	22
Haynes (2005)[19]	Websites, published filters	RCTs	21 [2]
Leeflang (2006)[1]	Database search	Diagnostic test accuracy studies	12
Manriquez (2008)[22]	Published filters	RCTs	2 [1]
McKibbon (2009)[13]	Database search, websites, published filters	RCTs	38
Montori (2005)[24]	Published filters	Systematic reviews	10 [4]
Ritchie (2007)[2]	Database search, contacted experts, published filters	Diagnostic test accuracy studies	23
Vincent (2003)[20]	Database search, websites	Diagnostic test accuracy studies	8 [3]
White (2001)[25]	Published filters	Systematic reviews	7 [5]
Whiting (2011)[18]	Contacted experts,	Diagnostic test	22

	database search, published filters	accuracy studies	
Wilczynski (2005)[21]	Published filters	Diagnostic test accuracy studies	4 [2]
Wong *et al*. (2006)[15]	Published filters	RCTs and systematic reviews	13

1

Gold standards

In search filter research a gold, or reference, standard is a set of relevant records against which the filter's performance can be assessed. For example, a collection of records of confirmed RCT studies would be used when testing the performance of a methodological search filter designed to identify RCTs.

Studies included in this review used a range of techniques to identify and/or create a gold standard against which to test the performance of multiple filters. One study did not use a gold standard.[16] Instead each of the filters was combined with single terms describing four topics (hypertension, hepatitis, diabetes and heart failure) and the retrieved studies were checked to confirm whether they were RCTs.

The size of gold standards used to test filter performance ranged from 33 to 1,955 records. None of the studies included in this review stated whether they had carried out a sample size calculation when developing their gold standard. (A sample size calculation is a statistical process that determines the minimum number of records required for a gold standard to provide accurate estimates of performance.) Four of the diagnostic test accuracy filter studies[2, 12, 18, 20] and 1 RCT filter study[23] limited their gold standard to specific clinical topics.

Ten studies developed their gold standard by hand-searching journals.[13, 15, 19, 21, 22, 24-28] The number of journals hand-searched ranged from 4 to 161. The time span covered by hand-searching varied from 1 to 23 years. All of the studies using hand-searching to create a gold standard had specific criteria for the identification of the desired study type for inclusion in their gold standard.

Of the ten studies identifying their gold standard from hand-searching journals, eight were studies where the authors had developed new search filters and then compared those filters to existing filters.[19, 21, 22, 24-28] One study that created a gold standard from hand-searching journals, created a "control set" of records from the same group of journals that were not the desired study design.[27]

Five studies developed a gold standard based on the studies included in systematic reviews (relative recall gold standard)[1, 2, 12, 18, 20] and four studies used database searches to identify records to include in their gold standard[12, 17, 23, 26]. The number of completed systematic reviews used as a source of gold standard records varied: one used included studies from 27 systematic reviews[1], one used included studies from 2 reviews[12], one used 7 reviews of diagnostic test accuracy studies[18] and a fourth used studies included in a single case study review[2]. One study which developed a diagnostic test accuracy study filter and compared it to published filters used the studies included in 16 reviews as their gold standard.[20]

Methods of testing

Four of the filter studies that used included studies from systematic reviews as their gold standard replicated the original searches where possible with the addition of the filters being tested.[1, 2, 12, 18] None of the original searches incorporated a study method search filter.[1, 2, 12, 18] A fifth study using references from systematic reviews as a gold standard combined the filters with "DVT terms" but did not specify what these terms were or if the original search strategy was used.[20]

The performance analyses carried out by Leeflang[1] and Ritchie[2] occurred after the original reviews (on which the gold standard was based) had been undertaken and therefore attempted to recreate a 'historical' search. Ritchie[2] noted a small discrepancy in the number of records retrieved between the original and re-run searches, while Leeflang[1], who could replicate only 6 out of 27 reviews, did not provide details of any differences in the number of retrieved records. Using the complete gold standard from the original reviews, Leeflang[1] tested whether those studies were captured by the filters being compared.

Two studies did not provide any information about whether the performance analysis had been undertaken concurrently with the reviews or at a later date.[12, 20] The Whiting review recreated the original subject search and compared using the subject search alone with using the subject search combined with 22 other filters.[18]

Four studies by the Hedges team at McMaster University used their internally developed database for testing filters, with the diagnostic test accuracy, RCT and systematic review subsets acting as gold standards.[13, 15, 19, 24] One of these studies did not undertake any new analysis, but collated the results from previous publications that had used a common gold standard.[15]

The economic filters study identified a gold standard by searching NHS EED.[17] Published MEDLINE and Embase economic filters were then tested on their ability to retrieve these gold standard records from MEDLINE and Embase. Corrao had no gold standard but manually checked whether the records retrieved after applying the filters were RCT studies.[16]

Studies that compared new search filters with existing filters can be divided into two groups based on the type of gold standard they used to compare filter performance. One group used a reference standard that had not been used to develop the new filter strategy so that all the filters in the comparison underwent external validation.[19, 23, 24, 27, 28] In other words, the performance of all the filters being compared was tested in a set of records that had <u>not</u> been used to develop any of the included filters. The other group of studies used the same reference standard that had been used in the development of the new filter, so while the new filter only underwent internal validation (filter performance was only tested on the one set of records which had also been used to develop the new filter) the comparison filters underwent external validation.[20-22, 25, 26] In the latter group, bias in favour of the new filters risks being introduced.

Translation of filters

Search filters were developed using a range of different search platforms (or interfaces) including PubMed, Ovid, or WebSPIRS for MEDLINE filters. Any study comparing the performance of filters may, therefore, need to "translate" the filters from the syntax used in the original development interface to the syntax required by the interface used in the filter comparison.

Four of the studies included in this review did not translate or adapt the filters they compared because the filters had been developed in the same interface as was used in the performance comparison.[15, 16, 22, 26] Where one or more filters required translation, most of the studies comparing performance of existing filters reported the complete details of the changes made so that the accuracy of the translation could be verified.[1, 12, 13, 17, 18] In contrast, most of the studies reporting the development of new filters that included a comparison with existing filters did not mention the requirement to translate any of the filters or provide details of the translation, so it is unclear if valid comparisons were being made.[19, 23, 24, 27, 28] The review of economic evaluation filters applied an exclusion strategy (animal studies and publication types such as letters and editorials which are unlikely to be economic evaluations) to filters being tested in MEDLINE and Embase.[17]

Performance measures reported

The most commonly reported performance measures in studies comparing the performance of search filters were sensitivity/recall and precision (Table 2). A total of 16 studies reported sensitivity/recall[2, 12, 13, 15, 17-28] and 13 studies reported precision values[2, 12, 13, 15-18, 21, 23-26, 28]. Specificity was reported in seven studies[13, 15, 19, 21, 22, 24, 27].

One study that did not use a gold standard, could not calculate sensitivity and instead reported the proportion of retrieved records that met the authors' criteria for being an RCT.[16] Another study calculated the proportion of gold standard records retrieved and missed for each filter.[1] Where the original search strategy could not be replicated this paper reported the number needed to read (NNR).[1] Bachman[28] reported the NNR for the filter they developed but not the previously published filter they used as a comparator. Whiting[18] reported NNR and the number of records missed from the reference set (gold standard).

No studies comparing filter performance reported accuracy values (proportion of records correctly retrieved or correctly not retrieved as a proportion of all records). The Manriquez[22] study reporting the development of an RCT filter for LILACS did report accuracy values, as did Wilczynski[21] for their diagnostic test accuracy study filters. Additional measures reported in performance comparisons were:

- number of records retrieved (NRR)[2];
- retrieval gain (absolute and percentage variations in number of citations retrieved)[16]
- proportion of articles missed per original review[1]
- proportion of articles not identified per year[1]

- diagnostic odds ratios (the ratio of the odds of the filter correctly identifying studies with the desired methodology)[27]
- number of relevant articles retrieved[26].

Confidence intervals surrounding performance results were reported by three studies that compared the performance of existing search filters.[13, 15, 18] Five of the studies comparing the performance of developed search filters with existing search filters reported confidence intervals.[21, 22, 24, 27, 28]

Table 2: Measures reported in filter performance comparisons

Performance measure	Study design being identified	Number of studies reporting the measure
Sensitivity/Recall	Economic evaluation	1
	Diagnostic test accuracy	7
	RCT	5
	Systematic review	4
Precision	Economic evaluation	1
	Diagnostic test accuracy	5
	RCT	4
	Systematic review	4
Specificity	Economic evaluation	0
	Diagnostic test accuracy	2
	RCT	4
	Systematic review	2
Accuracy	Economic evaluation	0
	Diagnostic test accuracy	1
	RCT	1
	Systematic review	0
NNR (number needed to read)	Economic evaluation	0
	Diagnostic test accuracy	3
	RCT	0
	Systematic review	1
Other (as detailed in text)	Economic evaluation	0
	Diagnostic test accuracy	4
	RCT	1
	Systematic review	1

Methods used to display performance comparisons/data

All of the studies included in the review displayed results using a table format, with only two studies supplementing tables of results with graphical (non-table) displays of comparative data.[1, 18] None of the studies reporting the development of new filters displayed comparative performance in a graphical format.[19-28]

The majority of tables presenting performance comparison data displayed the filters as rows and performance measures as columns (an example is provided in Table 3). Results in tables were given as percentages or proportions in all included studies. Within tables, authors generally listed filter results in descending order by the measure of interest, for example, decreasing sensitivity. Four studies reporting the development of a filter only included data on comparative performance in the text of the study report.[19, 22, 27, 28]

Tables that did not list filter results in descending order by the measure of interest instead arranged results by:

- the database in which filters were tested[15, 21]

- strategy type (sensitive strategy, specific strategy, optimized strategy)[15, 21]
- filter criteria (sensitive, accurate, etc)[1]
- filter alone compared to a clinical subject strategy[12]
- with and without the use of an exclusion strategy[17]
- by clinical topic considered in the performance testing[12, 16]
- subject search alone compared to the same subject search with each test filter[18]
- author or source of published filters[21, 24]
- descending order of cumulative precision or cumulative sensitivity[26]

Tables were also used to present information on number of studies retrieved[12] and the specificity, sensitivity and precision of single terms[15]. One study that reported highest precision combined with sensitivity greater than 69% showed the results of the filters meeting these criteria in a separate table.[2]

Table 3: Example of a filter performance comparison table as commonly presented in the literature

Filter	Number of records retrieved	Filter sensitivity	Filter precision
RCT filter A	NN	X%	Y%
RCT filter B	NN	X%	Y%
RCT filter C	NN	X%	Y%

Leeflang[1] used a bar graph to display the average proportion of retrieved and missed gold standard records per filter tested (Figure 1).

Figure 1: Bar chart displaying comparative performance of filters for diagnostic accuracy studies as published by Leeflang.[1]

Reprinted from Journal of Clinical Epidemiology, 59 (3), Leeflang MM, Scholten RJ, Rutjes AW, Reitsma JB, Bossuyt PM, Use of methodological search filters to identify diagnostic accuracy studies can lead to the omission of relevant studies, p234-240, Copyright (2006), with permission from Elsevier."

Whiting[18] presented the overall sensitivity and specificity of each filter tested in a forest plot, including confidence intervals (Figure 2).

Figure 2: Forest plot of overall sensitivity and precision for each filter in the Whiting study.[18]

Reprinted from Journal of Clinical Epidemiology, 64(6), Whiting P, Westwood M, Beyson R, Burke M, Sterne JAC, Glanville J, Inclusion of methodological filters in searches for diagnostic test accuracy studies misses relevant studies, p602-607, Copyright (2011), with permission from Elsevier.

Discussion

Eighteen published papers met the criteria for inclusion in this review. No numerical syntheses of filter performance comparisons were identified which may be due to the limited availability of performance comparison papers. The majority of included studies reported the development of one or more new filters, and compared performance against existing filters as an adjunct to the main research. This would seem to indicate a focus within filters research on development of new, "better" filters rather than comparison of performance across existing filters. However, the proliferation in search filters may make it more difficult for searchers to quickly select the most appropriate filter for their particular purpose. The development of increasingly effective filters and the transparent reporting of performance comparisons are important in demonstrating improvements in comparison to current methodological filters.

The number of comparisons of performance varies across study designs. A single study was identified that compared the performance of economic evaluation filters[17], whereas studies reporting on the performance of diagnostic test accuracy and RCT filters were much more common. As there are several specialist economics databases (NHS Economic Evaluation Database, Health Economic Evaluations Database, CEA Registry and the PEDE database) it may be that filters for the retrieval of economic evaluation studies are being given a lower research priority than filters for other study designs such as RCTs and diagnostic test accuracy studies.

Reporting methods of comparison

It was difficult to assess the reliability of the methods used in studies comparing the performance of multiple search filters because the size of the gold standard, the method of testing, the performance measures reported and the presentation of results varied greatly across studies. In addition, amongst studies that developed new filters, the methodological detail provided in comparing filter performance with existing filters was limited.

The description of methods used in studies reporting the development of new filters and those comparing only published filter performance differed. Those developing new filters focused their methods section on describing the selection and combination of terms for use in the new filters, with only minimal detail provided in the sections dedicated to describing the comparison of the new filter performance against existing filters. The comparison was often secondary to the main analysis and suffered from a lack of transparency. In contrast, studies where the focus was on comparing the performance of multiple existing filters the methods of identifying and testing the published filters included in the study tended to be reported more fully.

Many filter development studies did not clearly explain how they had identified filters for inclusion in performance testing. Not reporting how filters were identified and whether they were developed in the same interface used for testing could have implications for reliability and bias within the study. If studies do not report how the filters used in comparisons were identified, it is not possible to determine whether the filters were selected in an unbiased fashion or if they might have been preferentially selected to suit the test environment. In this review, studies reporting the development and testing of one or more filters all found that the

new filter performed better than the existing filters used as comparators. This makes it particularly important that studies clearly report how filters are selected and the comparison performed as otherwise this could be a sign of bias in the results.

Details about the “translation” of published filters for new interfaces were lacking in many filter development studies. Generally more detail about methods of “translation” was provided in studies which reported filter performance comparisons separately from the development of new filters. Combined with the lack of information on the original interface used in the development of published filters, the lack of “translation” details in many filter development studies makes it almost impossible to determine the accuracy of any alterations. As incorrect or imprecise translation of a filter is likely to impact on the results retrieved, the lack of methodological detail in filter performance comparison is cause for concern.[29]

Almost all of the included studies used a gold standard to test the comparative performance of developed and existing filters. This would seem to indicate that using a gold standard to test and compare filter performance is widely accepted in the filter research community. However, the size of the gold standard used varied widely from tens to thousands of records. It is possible that the size and content of the gold standard may have an impact on the performance measures recorded for a specific filter and so it would be helpful if researchers could justify their choice, by for example, reporting a sample size calculation.

Some of the studies included in the review used a single gold standard for both developing a new filter and comparing the new filter with published filters. This could potentially introduce performance bias in favour of the new filter as the new filter only undergoes internal validation whilst the comparator filters undergo external validation. In other words, the new filter is only tested against the set of records it was developed from, while the comparator filters are tested against a set of records that are different from the gold standard which was used to develop them. When a filter is tested against the same set of records from which it was developed, it is likely that the filter will perform better than it might in a different sample of records.

Reporting performance measures

Sensitivity and precision appear to be considered the most useful measures of filter performance since they are the most commonly reported measures in the literature. As the same performance measures were reported in studies developing new search filters and studies reporting the comparative performance of existing filters this is one area of methodological consistency between the two types of performance comparison study included in this review.

There is a suggestion from the small number of studies included in this review that there are some measures that are preferentially reported in diagnostic test accuracy study filters, for example, number needed to read (NNR). Similarly to the metric ‘number needed to treat (NNT)’, NNR reflects the number of retrieved records that need to be reviewed to identify a relevant study. By reporting the NNR, studies seek to make it easier for searchers to determine how effective a filter will be in reducing the number of irrelevant records retrieved and therefore the relative reduction in time needed to identify relevant studies for inclusion or full-text retrieval.

The methods used to present the results of filter performance comparisons were limited to tables and, in two studies, graphs. Tables were by far the most common method of reporting results from filter performance comparisons, perhaps reflecting the difficulties in presenting filter performance comparisons visually. Many of these tables were long and complicated making interpretation of the results and the selection of an appropriate filter challenging. In most cases it would not be easy to identify the most suitable filter without reading several studies, including tables, in detail. A lack of time and search filter expertise potentially compounds the problem of selecting an appropriate filter based on performance data as it is currently reported in the literature.

Of the two graphics used in the included studies to present results, a design similar to a forest plot (Figure 2) may prove attractive to searchers as it is a familiar format used in systematic reviews and meta-analyses. This design may also make it easier to identify visually the most precise, most sensitive and best balanced filter. A further exploration of methods for graphically presenting filter performance comparisons would be useful to both researchers involved in filter performance research and searchers needing to identify a suitable filter for their project. A separate element of the MRC-funded project of which this review is a part, explores this area of performance visualization.

Limitations of this review

There are a number of potential limitations to this review. Firstly, due to time constraints it was not possible to undertake a full systematic review. It was also not possible to review all filters for all study methods. However, the review was focused on study types which were felt to be the key study designs of current interest in evidence-based health research. Finally, research carried out on the performance of multiple search filters that has not yet been published or has only been presented at conferences was excluded from the review, possibly resulting in some alternative formats for the presentation of results being missed. However, conference abstracts would be likely to report even fewer methodological details than was presented in the full papers included in this review.

Suggestions for future research

From the results of this review the following are suggested as areas for future research:

- A review of measures reported and methods of presentation in methodological filter performance comparisons for study designs not included in this review
- Studies to explore alternative methods of displaying performance results from multiple methodological search filters
- Explorations of methods for numerical synthesis of the results of several filter performance comparisons

Conclusions

By considering which performance measures are reported in methodological search filter comparisons and how those measures are presented, rather than the actual results of the performance comparisons, this review has shown how search filter research is moving

towards more regular performance assessment both when offering new filters and when reviewing the performance of published filters. However, this review has also shown that efforts to assess comparative filter performance are hindered by confusing presentation of results and lack of methodological detail which impedes an assessment of bias even when the underlying research may have been of sound methodological quality. While the most commonly reported performance measures come as little surprise, this review has highlighted the potential for novel and innovative methods of presenting results from filter performance comparisons to aid in search filter selection. Hopefully the results of this review will encourage authors considering publishing a filter development or comparison study to give further thought to how to undertake their research and how to present their results to readers.

References

1. Leeflang, M.M., Scholten, R.J., Rutjes, A.W., Reitsma, J.B., & Bossuyt, P.M. Use of methodological search filters to identify diagnostic accuracy studies can lead to the omission of relevant studies. *J Clin Epidemiol* 2006;**59**:234-240.
2. Ritchie, G., Glanville, J., & Lefebvre, C. Do published search filters to identify diagnostic test accuracy studies perform adequately? *Health Info Libr J* 2007;**24**:188-192.
3. Glanville, J., Fleetwood, K., Yellowlees, A., Kaunelis, D., & Mensinkai, S. Development and testing of search filters to identify economic evaluations in Medline and Embase. Canadian Agency for Drugs and Technologies in Health (CADTH), Canada, 2009.
4. Glanville, J., & Paisley, S. Searching for cost-effectiveness decisions. In: Shemilt, I., & Mugford, M., (eds). *Evidence-based Economics*. Wiley-Blackwell, Oxford, 2009.
5. Deurenberg, R., Vlayen, J., Guillo, S., et al. Standardization of search methods for guideline development: an international survey of evidence-based guideline development groups. *Health Info Libr J* 2008;**25**:23-30.
6. Jenkins, M. Evaluation of methodological search filters: a review. *Health Info Libr J* 2004;**21**:148-163.
7. Boluyt, N., Tkosvold, L., Lefebvre, C., Klassen, T.P., & Offringa, M. The usefulness of systematic review search strategies in finding child health systematic reviews in Medline. *Arch Pediatr Adolesc Med* 2008;**162**:111-116.
8. Sampson, M., Zhang, L., Morrison, A., et al. An alternative to the hand searching gold standard: validating methodological search fitlers using relative recall. 2006.
9. Royle, P., & Milne, R. Literature searching for randomized controlled trials used in Cochrane reviews: rapid versus exhaustive searches. *Int J Technol Assess Health Care* 2003;**19**:591-603.
10. Bardia, A., Wahner-Roedler, D.L., Erwin, P.L., & Sood, A. Search strategies for retrieving complementary and alternative medicine clinical trials in oncology. *Integr Cancer Ther* 2006;**5**:202-205.
11. Kastner, M., Wilczynski, N.L., McKibbon, A.K., Garg, A.X., & Haynes, R.B. Diagnostic test systematic reviews: bibliographic search filters ("Clinical Queries") for diagnostic accuracy studies perform well. *J Clin Epidemiol* 2009;**62**:974-981.
12. Doust, J.A., Oietrzak, E., Sanders, S., & Glasziou, P.P. Identifying studies for systematic reviews of diagnostic tests was difficult due to the poor sensitivity and precision of methodologic filters and the lack of information in the abstract. *J Clin Epidemiol* 2005;**58**:444-449.

13. McKibbon, K.A., Wilczynski, N.L., & Haynes, R.B. Retrieving randomized controlled trials from Medline: a comparison of 38 published search filters. *Health Info Libr J* 2009;**26**:187-202.
14. Royle, P., & Waugh, N. A simplified search strategy for identifying randomised controlled trials for systematic reviews of health care interventions: a comparison with more exhaustive strategies. *BMC Med Res Methodol* 2005;**5**:23.
15. Wong, S.S., Wilczynski, N.L., & Haynes, R.B. Comparison of top-performing strategies for detecting clinically sound treatment studies and systematic reviews in MEDLINE and EMBASE. *J Med Libr Assoc* 2006;**94**:451-455.
16. Corrao, S., Colomba, D., Arnone, S., et al. Improving efficacy of Pubmed Clinical Queries for retrieving scientifically strong studies on treatment. *J Am Med Inform Assoc* 2006;**13**:485-487.
17. Glanville, J., Kaunelis, D., & Mensinkai, S. How well do search filters perform in identifying economic evaluations in MEDLINE and EMBASE. *Int J Technol Assess Health Care* 2009;**25**:522-529.
18. Whiting, P., Westwood, M., Beynon, R., et al. Inclusion of methodological filters in searches for diagnostic test accuracy studies misses relevant studies. *J Clin Epidemiol* 2011.
19. Haynes, R.B., & Wilczynski, N.L. Optimal search strategies for retrieving scientifically strong studies of diagnosis from Medline: analytical survey. *BMJ* 2005;**330**:1179.
20. Vincent, S., Greenley, S., & Beaven, O. Clinical Evidence diagnosis: developing a sensitive search strategy to retrieve diagnostic studies on deep vein thrombosis: a pragmatic approach. *Health Info Libr J* 2003;**20**:150-159.
21. Wilczynski, N.L., Haynes, R.B., & Hedges, T. EMBASE search strategies for identifying methodologically sound diagnostic studies for use by clinicians and researchers. *BMC Med* 2005;**3**:7.
22. Manriquez, J.J. A highly sensitive search strategy for clinical trials in Literatura Latino Americana e do Caribe em Ciencias da Saude (LILACS) was developed. *J Clin Epidemiol* 2008;**61**:407-411.
23. Glanville, J.M., Lefebvre, C., Miles, J.N.V., & Camosso-Stefinovic, J. How to identify randomized controlled trials in MEDLINE: ten years on. *J Med Libr Assoc* 2006;**94**:130-136.
24. Montori, V.M., Wilczynski, N.L., Morgan, D., & Haynes, R.B. Optimal search strategies for retrieving systematic reviews from Medline: analytical survey. *BMJ* 2005;**330**:68.
25. White, V.J., Glanville, J.M., Lefebvre, C., & Sheldon, T.A. A statistical approach to designing search filters to find systematic reviews: objectivity enhances accuracy. *J Info Sci* 2001;**27**:357-370.
26. Boynton, J., Glanville, J., McDaid, D., & Lefebvre, C. Identifying systematic reviews in MEDLINE: developing an objective approach to search strategy design. *J Info Sci* 1998;**24**:137-157.
27. Deville, W., Bezemer, P.D., & Bouter, L.M. Publications on diagnostic test evaluation in family medicine journals: an optimal search strategy. *J Clin Epidemiol* 2000;**53**:65-69.
28. Bachman, L.M., Coray, R., Estermann, P., & Ter Riet, G. Identifying diagnostic studies in MEDLINE: reducing the number needed to read. *J Am Med Inform Assoc* 2002;**9**:653-658.
29. Bradley, S.M. Examination of the clinical queries and systematic review "hedges" in EMBASE and MEDLINE. *Journal of the Canadian Health Libraries Association* 2010;**31**:27-37.

Appendix 1: excluded studies

Study	Reason for exclusion
Bardia et al. (2006)[10]	Study compares performance of filters for complementary and alternative medicine rather than clinical trials methodology.
Kastner et al. (2009)[11]	Study examines performance of the PubMed clinical query sensitive search filter for diagnostic studies in MEDLINE and Embase. This is a comparison of a single filter in two interfaces and not a comparison of performance of multiple filters.
Royle et al. (2005)[14]	Study did not test filters, it checked which CENTRAL records were not retrieved by Cochrane HSSS and which RCTs not in CENTRAL had random$ in the record.

Appendix 2: tables of included studies

A: Studies reporting on comparative performance of existing filters

Study	What study type was the filter designed to retrieve? (number of filters included)	Which database were filters tested in?	How were filters identified for comparison?	Filter translation (if any)	How was the gold standard (GS) used in comparisons developed? [size of gold standard]	Method used to compare filter performance	Performance measures reported per filter
Corrao et al. (2006)[16]	RCTs (2)	PubMed	PubMed Clinical Queries specific therapy filter and author's modified version: addition of term "randomised [Title/Abstract]"	Not required	No gold standard	Retrieved citations were "formally checked" to confirm RCT study design.	No. retrieved that were confirmed RCTs; Precision; Retrieval gain (absolute and percentage)
Doust et al. (2005)[12]	Diagnostic test accuracy studies (5)	MEDLINE (WebSpirs)	Published strategies for diagnostic test systematic reviews (no further details given)	Reports conversion from PubMed to MEDLINE WebSpirs for one filter. Reproduces terms used for all filters but does not discuss translation.	Included studies from 2 systematic reviews. Studies identified from MEDLINE search using Clinical Queries diagnostic filter and reference check. [53 records]	Filter terms, complete filter and filter plus original subject searches for reviews. Does not report date searched.	Sensitivity/recall; Precision
Glanville et	Economic	MEDLINE	Consulted	Strategies	Records coded	Filters run in	Sensitivity;

al. (2009)[17]	evaluations (MEDLINE 14, Embase 8)	(Ovid) and Embase (Ovid)	websites and experts	adapted for Ovid "as necessary" and reported in supplementary table.	as economic evaluations in NHS EED (2000, 2003, 2006) and indexed in MEDLINE or Embase. [MEDLINE 1,955 records] [Embase 1,873 records]	MEDLINE and Embase for same years as GS with and without exclusions (animal studies and publication types unlikely to yield economic evaluations)	Precision
Leeflang et al. (2006)[1]	Diagnostic test accuracy studies (12)	PubMed	MEDLINE, Embase and Cochrane Methodology Register searches. Where multiple filters reported selected highest sensitivity, highest specificity and highest accuracy filters according to original author.	Strategies adapted for PubMed. Translations reported in full.	Included studies from 27 systematic reviews. [820 records]	Filters run against PubMed records. Replicated original searches for 6 reviews with addition of filters and using same time frame.	NNR; Proportion of original articles missed; Average proportion of retrieved and missed GS records per filter (bar chart); Proportion of articles not identified per year (graph).
McKibbon et al. (2009)[13]	RCTs (38)	MEDLINE (Ovid)	Database (PubMed) searches, web searches, consulted websites,	Strategies translated for Ovid. Translated filters reported in Appendix.	Hand-searching 161 journals (2000). [1,587 records]	Filters run in Clinical Hedges database (49,028 Medline records from	Sensitivity/recall; Precision; Specificity; Confidence intervals.

			reviewed bibliographies, personal files			hand search journals).	
Ritchie et al. (2007)[2]	Diagnostic test accuracy studies (23)	MEDLINE (Ovid)	MEDLINE search, personal files, contacted experts	Reports one strategy translated from SilverPlatter to Ovid.	Included studies from one review indexed in MEDLINE. [160 records]	Replicated original review search (noted small discrepancy in results) with addition of filters	Sensitivity/recall; Precision; NRR
Whiting et al. (2011)[18]	Diagnostic test accuracy studies (22)	MEDLINE (Ovid)	MEDLINE (Ovid) search; consulted experts.	Details of translations to MEDLINE (Ovid) syntax reported as an Appendix.	References from 7 systematic reviews of diagnostic test accuracy studies that had not used methodological filters in the original search strategy. [506 records]	Compared performance of subject searches with that of filtered searches.	Sensitivity/recall; Precision; NNR; Number of missed records; Confidence intervals reported
Wong et al. (2006)[15]	MEDLINE RCTs (3) Embase RCTs (3) MEDLINE systematic reviews (3) EMBASE systematic	MEDLINE (Ovid) and Embase (Ovid)	Strategies developed by the authors and previously published	Not required	Hand-searching of 161 journals for MEDLINE and 55 for Embase. Not an external GS [RCT records: 930 MEDLINE, 1,256 Embase] [Systematic	None. Re-analysis comparing results of previous publications	Sensitivity/recall; Precision; Specificity; Confidence intervals.

	reviews (4)	review records: 753 MEDLINE 220 Embase]

NNR = Number needed to read; GS = gold standard; NRR = number of records retrieved.

B: Studies reporting on the development of one or more filters and their performance in comparison to previously published filters

Study	What study type was the filter designed to retrieve?	Which database were filters tested in?	How were filters identified for comparison?	Filter translation (if any)	How was the gold standard (GS) used in comparisons developed? [size of gold standard]	Method used to compare filter performance	Performance measures reported per filter
Bachman et al. (2002)[28]	Diagnostic test accuracy studies 1 developed (highest sensitivity x precision) 1 published (Haynes 1994)	MEDLINE (Datastar)	PubMed Clinical Query (Haynes 1994)	Does not discuss translation or reproduce Haynes strategy used.	Hand search of 4 journals from 1994 [53 records] and 4 different journals from 1999 [61 records].	External validation. Direct comparison of developed filter and current PubMed filter.	Sensitivity/recall; Precision; NNR (for developed filter only); Confidence intervals.
Boynton et al. (1998)[26]	Systematic reviews 11 developed 4 published	MEDLINE (Ovid)	Published strategies from Ovid MEDLINE	Translation not required.	Hand-searching of 6 journals from 1992 and 1995. [288 records]	Internal validation. Compared filter performance against a "quasi-gold standard".	Sensitivity/recall (described as cumulative); Precision (described as cumulative); Total articles retrieved; Number of relevant articles retrieved
Deville et al.	Diagnostic	MEDLINE	Only extensive	Not specified	Internal	Internal and	Internal

(2000)[27]	test accuracy studies <u>Internal validation:</u> 4 developed 1 published (Haynes 1994 sensitive strategy) <u>External validation:</u> 1 developed (most sensitive) 1 published (Haynes 1994 sensitive strategy)	(interface unknown)	paper on diagnostic filters (Haynes 1994)	but Haynes filter reproduced.	<u>validation set:</u> Hand search of 9 family medicine journals indexed in MEDLINE 1992-1995. Database search of MEDLINE 1992-1995 to create "control set". [75 records in gold standard, 137 records in "control set"] <u>External validation set:</u> 33 papers on physical diagnostic tests for meniscal lesions, no further details supplied.	external validation. Compared retrieval of published and developed strategies.	<u>validation:</u> Sensitivity/recall; Specificity; Diagnostic odds ratio; Confidence intervals <u>External validation:</u> Sensitivity/recall; Predictive value
Glanville et al. (2006)[23]	RCTs 6 developed 6 published	MEDLINE (Ovid)	Published strategies reporting over 90% sensitivity and with over 100 records in the gold	Not specified and filters not reproduced	Database search of MEDLINE (Ovid) 2003 using 4 clinical MeSH terms. Results	External validation. Compared retrieval in MEDLINE of 4 clinical MeSH	Sensitivity/recall; Precision

			standard used for development		assessed to identify indexed and non-indexed trials. [424 records]	terms with each comparator filter.	
Haynes et al. (2005)[19]	RCTs 2 developed (best sensitivity, best specificity) 19 published	MEDLINE (Ovid)	University filters website and known published papers. Selected strategies that had been tested against gold standards based on a hand search of published literature and for which MEDLINE records were available from 1990 onwards.	Not specified and filters not reproduced.	Hand-searching of 161 journals from 2000. [657 records]	External validation. Compared performance but full results not presented.	Sensitivity/recall; Specificity
Manriquez (2008)[22]	RCTs 1 developed 1 published (Castro 1999)	LILACS	Published filters	Not required (both developed and published filters designed for LILACS)	Hand searching of 44 journals published between 1981 and 2004 and indexed in LILACS. [267 records]	Internal validation. Compared ability to retrieve clinical trials included in the gold standard from the LILACs interface.	Sensitivity/recall; Specificity; Accuracy Confidence intervals

Montori et al. (2005)	Systematic reviews 4 developed 6 published	MEDLINE (Ovid)	"Most popular" published filters	Not specified and filters used not reproduced	Hand searching of 161 journals indexed in MEDLINE in 2000. [735 records]	External validation. Compared filters against validation standard.	Sensitivity/recall; Precision; Specificity; Confidence intervals
Vincent et al. (2003)[24]	Diagnostic test accuracy studies 3 developed 5 published	MEDLINE (Ovid)	Consulted websites; database search of MEDLINE	Not discussed but filters reproduced.	References from 16 systematic reviews. [126 records]	Internal validation. Compared sensitivity of developed and published strategies using reference set of MEDLINE records.	Sensitivity/recall
White et al. (2001)[25]	Systematic reviews 5 developed 2 published	MEDLINE (Ovid CD-ROM 1995-Sep 1998)	Published filters	Translated some filters from MEDLINE (Dialog) to MEDLINE (Ovid) syntax.	Hand searching of 5 journals from 1995 and 1997. Quasi gold standard of systematic reviews. [110 records]	Internal validation. Compared performance in "real world" search interface using quasi gold standard.	Sensitivity/recall; Precision
Wilczynski et al. (2005)[21]	Diagnostic test accuracy studies	Embase (Ovid)	Published filters	Not discussed but Bachman strategies	Hand searching of 55 journals from 2000. [97 records]	Internal validation. Compared performance of	Sensitivity/recall; Precision; Specificity; Accuracy; Confidence

	2 developed (most sensitive, most specific) 2 published (Bachman 2003 most sensitive and most specific)	reproduced.	developed and published filters in retrieving "methodologically sound" diagnostic studies.	intervals for differences between developed and published filters reported.

NNR = number needed to read