Cochrane Library

Cochrane Database of Systematic Reviews

Search strategies to identify diagnostic accuracy studies in MEDLINE and EMBASE (Review)

Beynon R, Leeflang MM, McDonald S, Eisinga A, Mitchell RL, Whiting P, Glanville JM

Beynon R, Leeflang MM, McDonald S, Eisinga A, Mitchell RL, Whiting P, Glanville JM.
Search strategies to identify diagnostic accuracy studies in MEDLINE and EMBASE.
Cochrane Database of Systematic Reviews 2013, Issue 9. Art. No.: MR000022.
DOI: 10.1002/14651858.MR000022.pub3.

www.cochranelibrary.com

WILEY

TABLE OF CONTENTS

[Methodology Review]

Search strategies to identify diagnostic accuracy studies in MEDLINE and EMBASE

Rebecca Beynon[1], Mariska M.G. Leeflang[2], Steve McDonald[3], Anne Eisinga[4], Ruth L Mitchell[5], Penny Whiting[6], Julie M Glanville[7]

[1]School of Social and Community Medicine, Bristol, UK. [2]Department of Clinical Epidemiology and Biostatistics, Academic Medical Center, J1B-207-1, AMSTERDAM, Netherlands. [3]School of Public Health & Preventive Medicine, Monash University, Melbourne, Australia. [4]UK Cochrane Centre, Oxford, UK. [5]Cochrane Renal Group, Centre for Kidney Research, The Children's Hospital at Westmead, Westmead, Australia. [6]Kleijnen Systematic Reviews, Escrick, UK. [7]York Health Economics Consortium, York, UK

Contact: Mariska M.G. Leeflang, Department of Clinical Epidemiology and Biostatistics, Academic Medical Center, J1B-207-1, P.O. Box 22700, AMSTERDAM, 1100 DE, Netherlands. m.m.leeflang@amc.uva.nl.

Editorial group: Cochrane Methodology Review Group.
Publication status and date: New, published in Issue 9, 2013.

Citation: Beynon R, Leeflang MM, McDonald S, Eisinga A, Mitchell RL, Whiting P, Glanville JM. Search strategies to identify diagnostic accuracy studies in MEDLINE and EMBASE. *Cochrane Database of Systematic Reviews* 2013, Issue 9. Art. No.: MR000022. DOI: 10.1002/14651858.MR000022.pub3.

ABSTRACT

Background

A systematic and extensive search for as many eligible studies as possible is essential in any systematic review. When searching for diagnostic test accuracy (DTA) studies in bibliographic databases, it is recommended that terms for disease (target condition) are combined with terms for the diagnostic test (index test). Researchers have developed methodological filters to try to increase the precision of these searches. These consist of text words and database indexing terms and would be added to the target condition and index test searches.

Efficiently identifying reports of DTA studies presents challenges because the methods are often not well reported in their titles and abstracts, suitable indexing terms may not be available and relevant indexing terms do not seem to be consistently assigned. A consequence of using search filters to identify records for diagnostic reviews is that relevant studies might be missed, while the number of irrelevant studies that need to be assessed may not be reduced. The current guidance for Cochrane DTA reviews recommends against the addition of a methodological search filter to target condition and index test search, as the only search approach.

Objectives

To systematically review empirical studies that report the development or evaluation, or both, of methodological search filters designed to retrieve DTA studies in MEDLINE and EMBASE.

Search methods

We searched MEDLINE (1950 to week 1 November 2012); EMBASE (1980 to 2012 Week 48); the Cochrane Methodology Register (Issue 3, 2012); ISI Web of Science (11 January 2013); PsycINFO (13 March 2013); Library and Information Science Abstracts (LISA) (31 May 2010); and Library, Information Science & Technology Abstracts (LISTA) (13 March 2013). We undertook citation searches on Web of Science, checked the reference lists of relevant studies, and searched the Search Filters Resource website of the InterTASC Information Specialists' Sub-Group (ISSG).

Selection criteria

Studies reporting the development or evaluation, or both, of a MEDLINE or EMBASE search filter aimed at retrieving DTA studies, which reported a measure of the filter's performance were eligible.

Data collection and analysis

The main outcome was a measure of filter performance, such as sensitivity or precision. We extracted data on the identification of the reference set (including the gold standard and, if used, the non-gold standard records), how the reference set was used and any limitations, the identification and combination of the search terms in the filters, internal and external validity testing, the number of filters evaluated, the date the study was conducted, the date the searches were completed, and the databases and search interfaces used. Where 2 x 2 data were available on filter performance, we used these to calculate sensitivity, specificity, precision and Number Needed to Read (NNR), and 95% confidence intervals (CIs). We compared the performance of a filter as reported by the original development study and any subsequent studies that evaluated the same filter.

Main results

Nineteen studies were included, reporting on 57 MEDLINE filters and 13 EMBASE filters. Thirty MEDLINE and four EMBASE filters were tested in an evaluation study where the performance of one or more filters was tested against one or more gold standards. The reported outcome measures varied. Some studies reported specificity as well as sensitivity if a reference set containing non-gold standard records in addition to gold standard records was used. In some cases, the original development study did not report any performance data on the filters. Original performance from the development study was not available for 17 filters that were subsequently tested in evaluation studies. All 19 studies reported the sensitivity of the filters that they developed or evaluated, nine studies reported the specificities and 14 studies reported the precision.

No filter which had original performance data from its development study, and was subsequently tested in an evaluation study, had what we defined a priori as acceptable sensitivity (> 90%) and precision (> 10%). In studies that developed MEDLINE filters that were evaluated in another study (n = 13), the sensitivity ranged from 55% to 100% (median 86%) and specificity from 73% to 98% (median 95%). Estimates of performance were lower in eight studies that evaluated the same 13 MEDLINE filters, with sensitivities ranging from 14% to 100% (median 73%) and specificities ranging from 15% to 96% (median 81%). Precision ranged from 1.1% to 40% (median 9.5%) in studies that developed MEDLINE filters and from 0.2% to 16.7% (median 4%) in studies that evaluated these filters. A similar range of specificities and precision were reported amongst the evaluation studies for MEDLINE filters without an original performance measure. Sensitivities ranged from 31% to 100% (median 71%), specificity ranged from 13% to 90% (median 55.5%) and precision from 1.0% to 11.0% (median 3.35%).

For the EMBASE filters, the original sensitivities reported in two development studies ranged from 74% to 100% (median 90%) for three filters, and precision ranged from 1.2% to 17.6% (median 3.7%). Evaluation studies of these filters had sensitivities from 72% to 97% (median 86%) and precision from 1.2% to 9% (median 3.7%). The performance of EMBASE search filters in development and evaluation studies were more alike than the performance of MEDLINE filters in development and evaluation studies. None of the EMBASE filters in either type of study had a sensitivity above 90% and precision above 10%.

Authors' conclusions

None of the current methodological filters designed to identify reports of primary DTA studies in MEDLINE or EMBASE combine sufficiently high sensitivity, required for systematic reviews, with a reasonable degree of precision. This finding supports the current recommendation in the *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy* that the combination of methodological filter search terms with terms for the index test and target condition should not be used as the only approach when conducting formal searches to inform systematic reviews of DTA.

PLAIN LANGUAGE SUMMARY

Search strategies to identify diagnostic accuracy studies in MEDLINE and EMBASE

A diagnostic test is any kind of medical test performed to help with the diagnosis or detection of a disease. A systematic review of a particular diagnostic test for a disease aims to bring together and assess all the available research evidence. Bibliographic databases are usually searched by combining terms for the disease with terms for the diagnostic test. However, depending on the topic area, the number of articles retrieved by such searches may be very large. Methodological filters consisting of text words and database indexing terms have been developed in the hope of improving the searches by increasing their precision when these filters are added to the search terms for the disease and diagnostic test. On the other hand, using filters to identify records for diagnostic reviews may miss relevant studies while at the same time not making a big difference to the number of studies that have to be assessed for inclusion. This review assessed the performance of 70 filters (reported in 19 studies) for identifying diagnostic studies in the two main bibliographic databases in health, MEDLINE and EMBASE. The results showed that search filters do not perform consistently, and should not be used as the only approach in formal searches to inform systematic reviews of diagnostic studies. None of the filters reached our minimum criteria of a sensitivity greater than 90% and a precision above 10%.

BACKGROUND

As with Cochrane reviews of interventions, Cochrane diagnostic test accuracy (DTA) reviews should aim to identify and evaluate as much available evidence about a specific topic as possible within the available resources (DeVet 2008). Thus, a systematic and extensive search for eligible studies is an essential step in any review. Recommendations for searching for DTA studies are that electronic bibliographic databases, such as MEDLINE and EMBASE, should be searched by combining search terms for disease indicators (target condition) with terms for the diagnostic test (index test) (DeVet 2008). Depending on the topic area, the number of articles retrieved by such searches may be too large to be processed with the available resources. A number of methodological filters consisting of text words and database specific indexing terms (such as MEDLINE Medical Subject Headings (MeSH)) have been developed in an attempt to increase the precision of searches and reduce the resources required to process results. These search filters are typically added to a search strategy consisting of the target condition and index test(s).

Methodological search filters have been developed for retrieving articles relating to many types of clinical question, including those about aetiology, diagnosis, prognosis and therapy. These filters are typically combinations of database indexing terms or text words, or both, that reflect the study design and statistical methods reported by the articles' authors. For example, Haynes and co-workers have developed a series of filters to assist searchers to retrieve articles according to aetiology, diagnosis, prognosis or therapy (Haynes 1994; Haynes 2005; Haynes 2005a; Wilczynski 2003; Wilczynski 2004). They are available as 'Clinical Queries' limits in both PubMed and via the OvidSP interfaces for MEDLINE and EMBASE (NLM 2005; OvidSP 2013; OvidSP 2013a).

Methodological search filters have proved to be particularly effective in identifying intervention (therapy) studies. Within The Cochrane Collaboration, a highly sensitive search strategy is widely used for identifying reports of randomised trials in MEDLINE (Lefebvre 2011).

For DTA studies, however, the relevant methodology is often not well reported by authors in their titles and abstracts. In addition, MEDLINE lacks a suitable publication type indexing term to apply to DTA studies. EMBASE has recently introduced a check tag for DTA studies (diagnostic test accuracy) but this is only being prospectively applied. Some relevant indexing terms do exist in both EMBASE and MEDLINE, for example sensitivity and specificity, however they are inconsistently assigned by indexers to DTA studies (Fielding 2002; Wilczynski 1995; Wilczynski 2005a; Wilczynski 2007). A consequence of adding filters to subject and index term strategies to identify records for DTA reviews is that relevant studies might be missed without, at the same time, significantly reducing the number of studies that have to be assessed for inclusion (Doust 2005; Leeflang 2006; Whiting 2008; Whiting 2011).

We conducted a methodology review of empirical studies that reported the development and evaluation of methodological search filters to retrieve reports of DTA studies in MEDLINE and EMBASE to assess the value of adding methodological search filters to search strategies to identify records for inclusion in DTA reviews. Until now, a comprehensive and systematic review of studies that develop or evaluate diagnostic search filters has not been published. The findings of this review will help to elucidate the performance of these filters to find studies relevant to diagnostic systematic reviews and to allow a recommendation for their use (or not) when conducting literature searches.

OBJECTIVES

To systematically review empirical studies that report the development or evaluation, or both, of methodological search filters designed to retrieve diagnostic test accuracy (DTA) studies in MEDLINE and EMBASE.

METHODS

Criteria for considering studies for this review

Types of studies

Primary studies of any design were included. Studies in which the main objective was the development or evaluation, or both, of a methodological filter for the purpose of searching for DTA studies in MEDLINE and EMBASE were eligible. We defined a development study as one in which a new filter was conceived, tested in a reference set of diagnostic studies, and the performance reported. An evaluation study was one in which a filter from a development study publication was tested in a new reference set and the performance reported. A study could be both a development and an evaluation study if it reported on the development and performance of a newly designed filter and evaluated a filter which had previously been published by a different development study. We also included filters assessed in evaluation studies for which there was no corresponding development study publication. We excluded studies that developed or evaluated filters designed to retrieve clinical prediction studies or prognostic studies.

Types of data

Eligible studies must have reported the performance of search filters using a recognised measure, such as sensitivity or precision.

Types of methods

Assessments of the performance of search strategies for identifying reports of DTA in MEDLINE and EMBASE.

Types of outcome measures

Eligible outcome measures were those that assessed the accuracy of the search.

Primary outcomes

Measures of search performance, including:

- sensitivity (proportion of relevant reports correctly retrieved by the filter);
- specificity (proportion of irrelevant reports correctly not retrieved by the filter);
- accuracy (the highest possible sensitivity in combination with the highest possible specificity);
- precision (the number of relevant reports retrieved divided by the total number of records retrieved by the filter).

We defined a priori the levels of sensitivity (> 90%) and precision (> 10%) from the external validation of evaluation studies as the acceptable threshold for use when searching for DTA studies.

Secondary outcomes

- Number Needed to Read (NNR) (also called Number Needed to Screen), which is the inverse of the precision (Bachmann 2002).

Search methods for identification of studies

Electronic searches

The following databases were searched to identify relevant studies: MEDLINE (1950 to week 1 November 2012); EMBASE (1980 to 2012 Week 48); the Cochrane Methodology Register (Issue 3, 2012); ISI Web of Science (11 January 2013); PsycINFO (13 March 2013); Library and Information Science Abstracts (LISA) (31 May 2010); and Library, Information Science & Technology Abstracts (LISTA) (13 March 2013). Three information specialists developed and conducted the searches. The search strategies are listed in the appendices (Appendix 1; Appendix 2; Appendix 3; Appendix 4; Appendix 5; Appendix 6; Appendix 7). No language restrictions were applied.

Searching other resources

We also undertook citation searches of the included studies on Web of Science. Furthermore, reference lists of all relevant studies were assessed (Horsley 2011) and the Search Filters Resource website of the InterTASC Information Specialists' Sub-Group (ISSG) was screened (InterTASC 2011). InterTASC is a collaboration of six academic units in the UK who conduct and critique systematic reviews for the National Institute for Health and Care Excellence.

Data collection and analysis

Selection of studies

Two authors independently screened the titles and abstracts of all retrieved records. Inclusion assessment of full papers was conducted by one author and checked by a second. Any disagreements were resolved through discussion or referral to a third author.

Data extraction and management

Data extraction was performed by one author and checked by a second; disagreements were resolved through discussion. The ISSG Search Filter Appraisal Checklist (Glanville 2008) was used to structure the data extraction and assessment of methodological quality. This checklist was developed using consensus methods and tested on several filters. It assesses the scope of the filter (limitations, generalisability and obsolescence), and the methods used to develop the filter, including the generation of the reference set.

Data were extracted on the characteristics of the reference set (inclusion of gold and non-gold standard records, years of publication of the records, journals covered, inclusion criteria, size); how search terms were identified; presence of internal and external validity testing; and any limitations or comparisons between studies. In the context of filter development, the reference set is the same as the reference standard or gold standard in DTA studies. In contrast, the gold standard in the context of filter development is equivalent to diseased individuals in diagnostic accuracy studies (that is the 'relevant' studies) and the non-gold standard is equivalent to non-diseased individuals (that is the non-relevant studies).

Data were also extracted on the date the study was conducted; the date the searches were completed; the database(s) and search interface(s) used; the outcome measures of performance (sensitivity, specificity, precision) and their definitions; and whether the search strategy was developed for specific clinical areas or to identify diagnostic studies over a broad range of topics. We assessed whether the search strategies were described in sufficient detail to be reproducible (that is were the search terms and their combination reported, were the dates of the search reported, and was the interface and database reported?).

Where studies reported data on multiple filters, results were extracted for each filter. However, for filter development studies, if data were also presented on the sensitivity and precision of all tested individual terms, only single term filters that the original authors selected as reporting best performance were extracted, as well as all multiple term filters.

Assessment of risk of bias in included studies

Bias occurs if systematic flaws or limitations in the design or conduct of a study distort the results. Applicability refers to the generalisability of results: can the results of the filter development or evaluation study be applied to other settings with different populations, index tests, reference standards or target conditions?

We identified three areas that we considered to have the potential to introduce bias or affect the applicability of the included studies.

1. Absence of DTA search strategy in reference set development: bias may be introduced when either a development or an evaluation study used a systematic review (or reviews) to provide studies for the reference set, and this systematic review used a search strategy containing diagnostic terms to find primary studies. This could introduce bias because the performance of a filter tested in this reference set will naturally be higher when the difficult to retrieve studies have been missed by the reference set search.

2. Choice of gold standard: concerns about applicability may be introduced in both development and evaluation studies in the generalisability of the filter to all diagnostic studies. Some filters have been developed or evaluated using a reference set that is composed of topic specific studies (such as studies on the diagnosis of deep vein thrombosis), whereas other reference sets will be generic (studies covering a wide range of diagnostic tests and conditions). Ideally, a filter will perform equally well across different topic areas but if it is only evaluated in one specific topic area its performance in other areas will be unclear.

3. Validation of filters in development studies: the process of validation can be split into two parts; the method of internal validation can have bias issues, while the method of external validation (if done) can have both applicability and bias issues. Internal validity is the ability of the filter to find studies from the reference set from which it was developed. A study could be at risk of bias if the internal validation set contained the references from which the filter terms were derived. External validity is the ability of the filter to find studies in a real-world setting (that is using a reference set composed of topic specific studies). This relates to how generalisable the results are to searching for diagnostic studies for different systematic review topics and most closely relates to how the filters would be used in practice by systematic reviewers. This issue only applies to development studies. A study which has

used external validation in a real-world setting will be judged to have low levels of concern about applicability. However, a study that includes external validity testing could still be at risk of bias if the validity testing occurred in a validation set containing the references used to derive the terms.

Data synthesis

We synthesised performance measures of the filters separately for MEDLINE and EMBASE. We tabulated the performance measures reported by development and evaluation studies grouped by individual filters, so that a comparison could be made between the original reported performance of a filter and its performance in subsequent evaluation studies. If sensitivity, specificity or precision together with 95% confidence intervals (CIs) were not reported in the original reports, these were calculated from the 2 x 2 data, where possible.

Each of the performance measures can be calculated as shown by the formulae below (a further description of performance measures is available in Appendix 8).

		Reference set	
		Gold standard records	**Non-gold standard records**
Searches incorporating methodological filter	Detected	a (true positive)	b (false positive)
	Not detected	c (false negative)	d (true negative)

Sensitivity = a/(a + c)

Precision = a/(a + b)

Specificity = d/(b + d)

Accuracy = (a + d)/(a + b + c + d)

Number needed to read = 1/(a/(a + b))

Reference set = gold standard + non-gold standard records = (a + b + c + d)

Gold standard = relevant DTA studies = a + c

NB. This is different to the gold (reference) standard in DTA studies, which is equivalent to the reference set in filter evaluations. The gold standard in DTA studies is able to correctly identify the true positives and as well as the true negatives, unlike the gold standard in a filter evaluation study which is limited to the true positives.

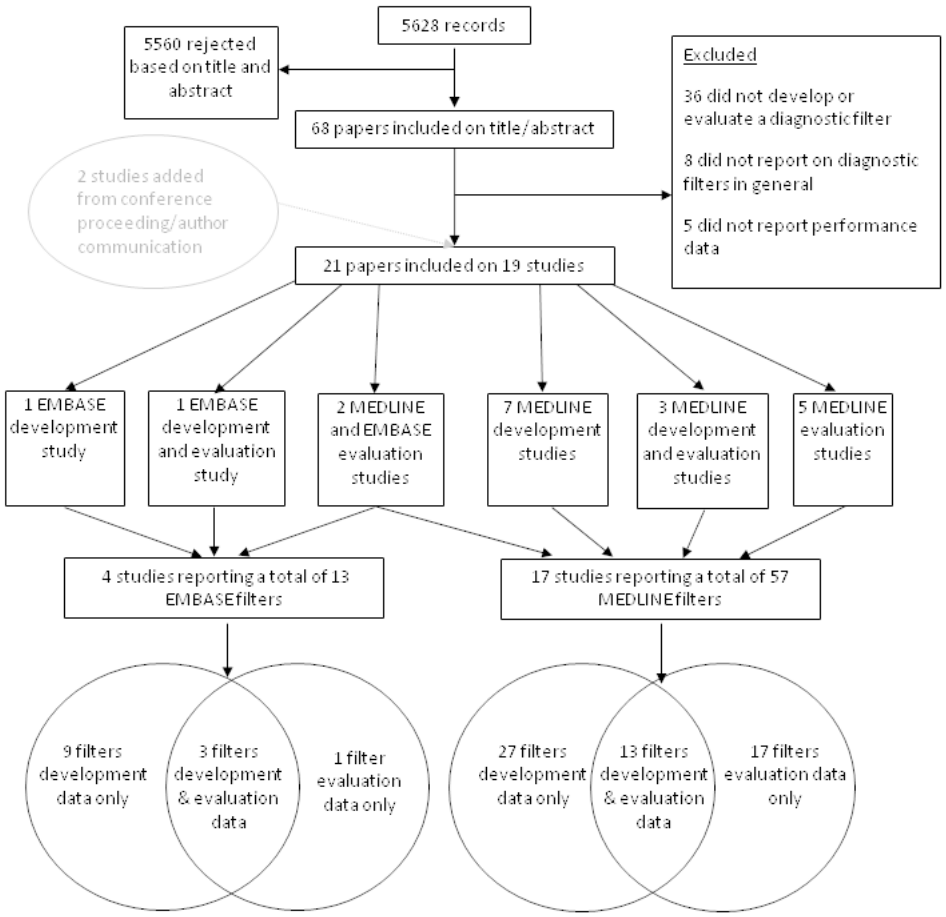
Paired results of either sensitivity and specificity or sensitivity and precision for each filter were displayed in receiver operating characteristic (ROC) plots. The original individual filter performance estimates from the development studies were plotted in the same ROC space as the individual filter performance estimates from the evaluation studies, to allow for visual inspection of disparities and similarities. We did not pool data due to heterogeneity across studies.

RESULTS

Description of studies

The searches retrieved 5628 records, of which 19 studies reported in 21 papers met the inclusion criteria (Figure 1). These assessed 57 MEDLINE filters and 13 EMBASE filters.

Figure 1. Study selection process.



MEDLINE search filters

Description of development studies

Ten studies reported on the development of 40 MEDLINE filters (range 1 to 12 filters per study). Key features of each study are summarized in the Characteristics of included studies table and Table 1. Thirty-one filters were composed of multiple terms and nine filters were single term strategies. Nine filters consisted of MeSH terms only, six filters had text words only, and 25 filters combined MeSH with text words. Full details of methods used in each study and the size of the reference set are given in Table 2. A description of each filter and its performance are listed in Table 3.

Method of identification of reference set records

Different methods were used to compile the reference sets. Six studies handsearched journals to obtain a database of 'gold standard' references reporting relevant DTA studies (Astin 2008; Berg 2005; Deville 2000; Haynes 1994; Haynes 2004).

Three studies used a relative recall reference standard, that is the reference set was based on studies included in systematic reviews. Deville 2002 used references from two published systematic reviews (on diagnosing knee lesions and the accuracy of urine dipstick testing) that had formed part of the first author's thesis. Noel-Storr 2011 used the references from a systematic review on the volume of evidence in biomarker studies in people with mild cognitive impairment. Another study (van der Weijden 1997) developed a reference set based on a personal literature database on erythrocyte sedimentation rate as a diagnostic test, compiled over 10 years 'by every means of literature searching'. Finally, one study used a validated filter to locate systematic reviews indexed in MEDLINE and EMBASE reporting on diagnostic tests for deep vein thrombosis, and used the studies included in these reviews as the reference set (Vincent 2003).

Two of the 10 studies described above included all articles that were retrieved by the search for gold standard records but which were subsequently rejected from the gold standard as the non-gold standard records in the reference set (Berg 2005; Noel-Storr 2011). A third study used the false positive articles selected by a search using a previously published diagnostic search strategy as the non-gold standard records in the reference set (Deville 2000). This study further restricted the non-gold standard studies by excluding reviews, meta-analyses, comments, editorials and animal studies. The remaining studies that included non-gold standard records in their reference set did not provide details on how these were identified.

Composition of reference set

Seven studies included both gold and non-gold standard references in their reference sets (Astin 2008; Bachmann 2002; Berg 2005; Deville 2000; Haynes 1994; Haynes 2004; Noel-Storr 2011) and two studies used only gold standard studies (van der Weijden 1997; Vincent 2003). One study did not give any details on the composition of the reference set (Deville 2002). It was possible to calculate sensitivity, specificity, precision and NNR from the studies that had a reference set compiled of both included DTA studies (gold standard references) and studies that did not meet the criteria of a DTA study (non-gold standard references) if 2 x 2 data were available. However, it was not possible to calculate specificity or precision from a reference set composed of only included DTA references. This was because the percentage of correctly non-identified studies cannot be calculated since data for only half of a 2 x 2 table were available.

Of the six studies that used handsearching to develop the reference set, two studies concentrated on specific topic areas. Astin 2008 included records on imaging as a diagnostic test and Berg 2005 included articles from the nursing literature on cancer-related fatigue diagnosis. The remaining studies that had a handsearched reference set were not topic specific. The two studies that used published systematic reviews to compile the reference set, and the study which used a personal literature database, were all topic specific.

Where reported, the mean number of gold standard studies in the reference set was 128 (range 33 to 333) from a mean of 35 journals (range 9 to 161). Of the studies that used reference sets which included non-gold standard as well as gold standard records, the mean number of overall references included was 8582 (range 238 to 48,881).

Method of identification of search terms

Three studies used the reference set to derive search terms by performing statistical analysis on terms found in titles, abstracts and subject headings (Astin 2008; Bachmann 2002; Deville 2000). Three studies adapted existing search strategies (Berg 2005; Deville 2002; Vincent 2003), one of which expanded the existing filters by adding frequently occurring MeSH terms and text words found in titles and abstracts of the reference set (Berg 2005). Vincent 2003 also combined the use of existing filters with the results of reference set analysis. Of the remaining four studies, one used expert knowledge of the field to generate a list of terms (Haynes 1994), one used expert knowledge and analysis of the reference set (Haynes 2004), one checked key publications for the definitions and terms used (van der Weijden 1997), and one analysed terms in 10 studies missed by the three most sensitive published filters (Noel-Storr 2011).

Description of studies that evaluated published MEDLINE filters

Ten evaluation studies that assessed 30 MEDLINE filters were included (Table 4; Table 5). Of these, three were development studies that also evaluated published filters and were therefore classed as both development and evaluation studies (Deville 2000; Noel-Storr 2011; Vincent 2003). Most filters (n = 23) were evaluated by at least two studies. The median number of filters evaluated in a each study was 6, but ranged from 1 (Deville 2000; Kastner 2009) to 22 (Noel-Storr 2011; Ritchie 2007; Whiting 2010).

Method of identification of reference set records

Seven studies used a relative recall reference set consisting of studies included in DTA systematic reviews (Doust 2005; Kastner 2009; Leeflang 2006; Noel-Storr 2011; Ritchie 2007; Vincent 2003; Whiting 2010). Of these, three studies located systematic reviews through electronic searches (Kastner 2009; Leeflang 2006; Vincent 2003) and four studies used a convenience sample of systematic reviews that either the authors or colleagues had undertaken themselves (Doust 2005; Noel-Storr 2011; Ritchie 2007; Whiting 2010). One study used references located through handsearching of the nine highest ranking journals available on MEDLINE (Deville 2000); one study handsearched three high ranking renal journals (as identified by the authors) for primary studies on the diagnosis of renal disease (Mitchell 2005); and one study used an electronic search for primary DTA studies related to venous thrombosis, venography and ultrasonography (Kassai 2006).

Three of the studies that used a relative recall reference set included reviews which used a methodological filter to find diagnostic studies in addition to terms for test and condition (Doust 2005; Kastner 2009; Vincent 2003). One of these studies supplemented the search, which had first used the Clinical Queries diagnostic filter in PubMed, by searching the reference lists of included studies (Doust 2005).

Two studies included all articles that were retrieved by the search for gold standard records but which were subsequently rejected from the gold standard as the non-gold standard records in the reference set (Kassai 2006; Ritchie 2007). A third study used the false positive articles selected by a search using a previously published diagnostic search strategy as the non-gold standard records in the reference set (Deville 2000). This study further restricted the non-gold standard studies by excluding reviews, meta-analyses, comments, editorials and animal studies. The remaining studies that included non-gold standard records in their reference set did not provide details on how these were identified.

Composition of reference set

Three of the seven studies derived their reference set from a systematic review that used gold standard and non-gold standard studies (Noel-Storr 2011; Ritchie 2007; Whiting 2010); the remaining four studies used a reference set comprised of only gold standard studies (Doust 2005; Kastner 2009; Leeflang 2006; Vincent 2003). The three studies which used an electronic search or a handsearch to find primary studies also included non-gold standard studies in their reference sets (Deville 2000; Kassai 2006; Mitchell 2005).

The number of gold standard studies included in the reference standard ranged from 53 from two systematic reviews (Doust 2005) to 820 from 27 reviews (Leeflang 2006). In all studies that also included non-gold standard studies, the number of irrelevant studies ranged from 1236 to 27,804.

Description of evaluated filters

All but one of the search strategies combined MeSH terms and text words; one used the single term strategy “specificity.tw” (Whiting 2010). Two of the evaluated filters that were displayed were based on the same original strategy by Haynes 1994. Falck-Ytter 2004 presented an alternative interpretation of the original filter in a PubMed format.

EMBASE search filters

Description of development studies

Two studies reported the development of 12 search filters for finding DTA studies indexed in EMBASE (Table 6; Table 7) (Bachmann 2003; Wilczynski 2005). Eleven of the filters were composed of multiple terms. Table 6 gives a summary of the study design characteristics of the included studies.

Method of identification of reference set records

In both studies the reference set was generated by handsearching journals, and included both gold standard and non-gold standard records. One study reported that the non-gold standard records were identified as all articles retrieved by the search that were not classified as gold-standard records (Bachmann 2003). The other study was not clear about how non-gold standard records were selected (Wilczynski 2005).

Composition of reference set

Both studies included both gold standard and non-gold standard records in the reference set.

Method of identification of search terms

One study used the reference set to derive filter terms using word frequency analysis (Bachmann 2003). The other study initially identified terms for the filter by consulting experts and then entered the terms into a logistic regression model to find the most frequently occurring terms (Wilczynski 2005).

Description of studies that evaluated published EMBASE filters

Three studies evaluated four filters designed to find DTA studies in EMBASE (Table 8; Table 9) (Kastner 2009; Mitchell 2005; Wilczynski 2005). One filter was evaluated by two studies, and three filters were evaluated by only one study. A summary of the study design characteristics of included studies is in Table 8.

Method of identification of reference set records

One study used studies from 12 published systematic reviews to construct the reference standard (Kastner 2009). The other two EMBASE filter studies identified primary DTA studies through handsearching (Mitchell 2005; Wilczynski 2005). Neither study which had included non-gold standard records described how those articles were identified.

Composition of reference set

Two studies included both gold standard and non-gold standard records in the reference set (Mitchell 2005; Wilczynski 2005). The number of gold standard records ranged from 96 to 441. The number of non-gold standard records ranged from 3984 to 27,575.

Description of evaluated filters

One evaluated filter consisted of MeSH terms and text words, the other three filters consisted of text words only. Every filter combined multiple terms.

Risk of bias in included studies

The methodological quality of the identified studies was not formally assessed using a validated tool, but we identified three areas that could affect the methodological quality of the studies in terms of the risk of bias and applicability as described above (see Assessment of risk of bias in included studies).

1. Use of systematic reviews to compile reference set search strategy

MEDLINE development and evaluation studies

Of the eight studies which used systematic reviews to compile their reference sets, three used reviews which did not include diagnostic terms in their search strategies and were at low risk of bias; one development and evaluation study and two evaluation studies specified that they only included systematic reviews which had not used a diagnostic search filter (Noel-Storr 2011; Ritchie 2007; Whiting 2010). The systematic reviews used by Whiting and Noel-Storr were conducted by the authors, therefore the reviewers could be sure that no such filter was applied. Ritchie also used a systematic review carried out by Whiting, which did not use a diagnostic filter.

Three studies used reviews with diagnostic terms in their search strategies and were therefore at high risk of bias. One was a development and evaluation study which contained the references from 16 systematic reviews and, of these, at least one used a diagnostic filter (Vincent 2003). Some of the other systematic reviews did not report whether they used a diagnostic filter or not, while the remaining studies were not available. Two evaluation studies also used reviews with diagnostic filter terms. Kastner's reference set contained the studies from 12 systematic reviews and, of these, just over half used diagnostic terms in their search strategies (Kastner 2009). Doust 2005 conducted two systematic reviews which were used in reference set development, and the search strategy for these applied the PubMed Clinical Queries filter for diagnostic studies.

For one development and one evaluation study, it was not clear whether the systematic reviews used a diagnostic filter in their searches (Deville 2002; Leeflang 2006). The risk of bias for these studies was unclear. The original source of the review used by Deville was not available (from the author's thesis), but a meta-analysis published by the same author on the same topic did describe the use of diagnostic terms in the search strategy. Leeflang stated in their discussion that while they attempted to exclude any review which used a diagnostic filter in their literature search, they found that of the 27 reviews where the studies were included, seven did not describe their search in detail.

EMBASE development and evaluation studies

Only one evaluation study, reporting an EMBASE filter, used the studies from systematic reviews to compile the reference set, and just over half of the 12 systematic reviews used diagnostic terms in their search strategies (Kastner 2009). This study was, therefore, judged to be at high risk of bias.

2. Choice of gold standard records

MEDLINE development and evaluation studies

Of 17 studies, three development and three evaluation studies used generic gold standard records and caused a low level of concern regarding applicability (Bachmann 2002; Haynes 1994; Haynes 2004; Kastner 2009; Leeflang 2006; Whiting 2010). Of these, the development studies handsearched a broad range of general medical journals while the evaluation studies used the included

studies from systematic reviews covering a range of diagnostic tests and conditions.

Four development studies used topic specific gold standard records to develop their filters (Astin 2008; Berg 2005; Deville 2002; van der Weijden 1997). In addition, the three studies which both developed and evaluated filters also used topic specific records (Deville 2000; Noel-Storr 2011; Vincent 2003). Four evaluation studies used topic specific gold standard records to test the performance of published filters (Doust 2005; Kassai 2006; Mitchell 2005; Ritchie 2007). These studies caused high levels of concern regarding applicability as they were only likely to be applicable to the particular topic area in which they were developed or evaluated. The topics included in these studies varied in their breadth, for example a very narrow topic was used by Kassai 2006 (limited to studies comparing ultrasound to venography for the diagnosis of deep vein thrombosis), whereas Deville 2000 included studies on diagnostic tests from nine family medicine journals. Other topics included diagnostic tests in radiology and biomarkers for mild cognitive impairment. Noel-Storr 2011 designed their filter to specifically retrieve longitudinal DTA studies and evaluated published filters for their ability to retrieve delayed cross-sectional DTA studies.

EMBASE development and evaluation studies

All but one of the four studies that developed or evaluated a diagnostic EMBASE filter used a set of gold standard records derived from on a broad range of topics and tests. One evaluation study handsearched the three top ranking renal journals for studies on the diagnosis of kidney disease (Mitchell 2005).

3. Validation of filters

MEDLINE development studies

Of the 10 studies reporting the development of a MEDLINE filter, two studies used discrete derivation and validation sets of references to test internal validity and were considered to be at low risk of bias (Astin 2008; Bachmann 2002). Astin handsearched six high ranking radiology journals to find studies for the derivation set and used a different set of six journals to compile studies for the validation set. Bachmann handsearched journals in different years; the studies found in 1989 comprised the set of references used to derive terms, while the studies from 1994 comprised the validation set.

Six of the remaining studies used an internal validation set which contained the references used to derive the terms for the filter and the studies were therefore judged to be at high risk of bias (Berg 2005; Deville 2000; Haynes 1994; Haynes 2004; Noel-Storr 2011; Vincent 2003). Of these studies, three independently selected terms to use as part of their filters, but the final strategies (made up of those terms) were derived from testing in the same set of references (Haynes 1994; Haynes 2004; Vincent 2003). Also of note, Noel-Storr 2011 derived filter terms by running published search filters in MEDLINE combined with a subject search, locating 10 papers that all filters missed and choosing a term from the title, abstract or keywords of each. These 10 papers were included in the reference set of 144 studies.

Two studies did not perform internal validity testing of the two filters that had been developed, rather specific diagnostic topics (reviews) were used only to externally validate (Deville 2002; van der Weijden 1997). These studies reported sensitivities > 90% for their most sensitive filters.

Four studies carried out external validation of their filters in a validation set that represented real-world settings, and the filters were judged to cause low levels of concern about applicability (Bachmann 2002; Deville 2000; Deville 2002; van der Weijden 1997). The remaining studies did not validate their filters in real-world settings and were considered to cause high levels of concern regarding applicability (Astin 2008; Berg 2005; Haynes 1994; Haynes 2004; Noel-Storr 2011; Vincent 2003).

EMBASE development studies

Both EMBASE development studies were at high risk of bias in this domain because neither study used a set of records independent from those used to derive the terms to internally validate their strategies (Bachmann 2003; Wilczynski 2005). Bachmann used word frequency analysis of all the titles and abstracts of studies included in the reference set to find and combine the 10 terms with the highest sensitivity and precision. Wilczynski first derived a list of potential diagnostic terms from clinical studies and then from clinicians and librarians. The individual search terms with sensitivity > 25% and specificity > 75%, when tested in the reference set, were then combined into the search strategies.

Neither study externally validated their newly developed filters and were therefore judged to have high concerns regarding applicability in this domain.

Effect of methods

1. Performance of MEDLINE filters as reported in development studies

Sensitivity ranged from 16% to 100% (median 86%; 39 filters, 10 studies), specificity ranged from 38% to 99% (median 88.5%; 30 filters, 6 studies) and precision ranged from 0.8% to 90% (median 9.3%; 32 filters, 8 studies) (Table 3).

2. Performance of evaluated MEDLINE filters

Performance data on each evaluated filter can be found in Table 10 and full search strategies can be found in Appendix 9. Thirteen of the 30 MEDLINE filters assessed by the evaluation studies had original performance data available from development studies. The other 17 filters were reported without any details on how they were developed or their performance.

None of the filters tested in development or evaluation studies had sensitivity > 90% and precision > 10%. The original studies reported sensitivities ranging from 55% to 100% (median 86%); evaluation studies reporting on the same 13 filters had sensitivities ranging from 14% to 100% (median 73%). Doust 2005 evaluated the two strategies with 100% sensitivity in a reference set composed of included studies from a systematic review of natriuretic peptides. The original searches for the two systematic reviews used the PubMed Clinical Queries filter (from Haynes 2004), supplemented by screening the reference lists of included studies. This might explain why the evaluated filters performed so well in this reference set. The sensitivities of the 18 evaluated filters that did not have accompanying original performance data ranged from 40% to 100% (median 71%).

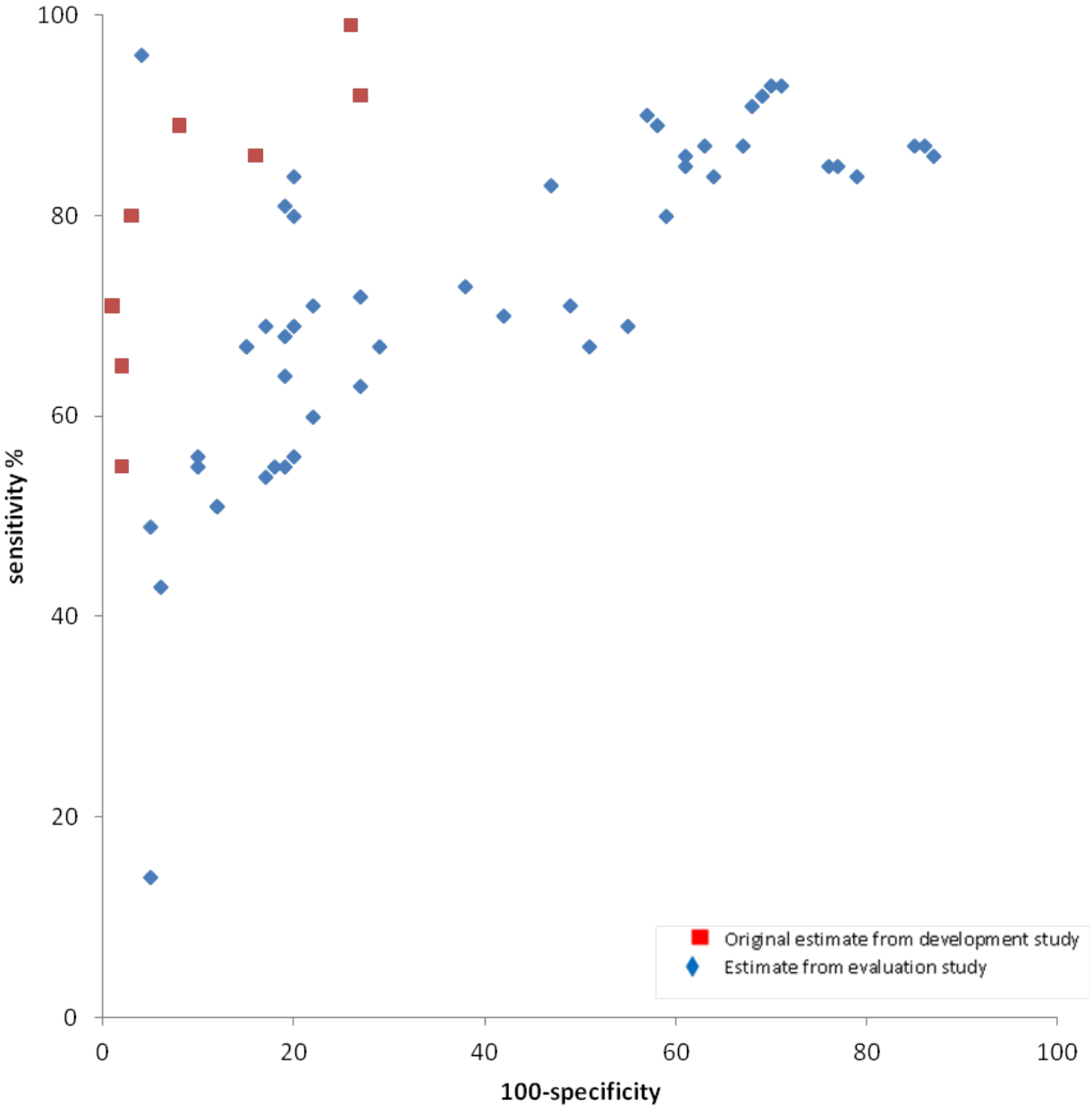
Specificity was only reported in the original study and three evaluation studies (Mitchell 2005; Noel-Storr 2011; Whiting 2010) for four filters and ranged from 73% to 98% (median 94.5%) in the original study and from 15% to 96% (median 81%) in the

evaluation studies. Similarly, precision was only reported in both the original study and evaluation studies for seven filters and ranged from 1.1% to 40% (median 9.5%) in the original study and from 0.2% to 16.7% (median 4%) in the evaluation studies. Similar ranges of specificities and precision were reported in the evaluation studies for the 17 filters without an original performance measure. Sensitivities ranged from 31% to 100% (median 71%), specificity ranged from 13% to 90% (median 55.5%) and precision from 1.0% to 11.0% (median 3.35%).

Original estimates of sensitivity were higher than those reported in the evaluation studies in 43 of 53 comparisons. (If an evaluation study had two reference sets, it contributed twice to the total number of comparisons for each filter evaluated.) Original estimates of specificity were higher in 10 of 14 comparisons, and precision was higher in 16 of 25 comparisons. None of the evaluated filters performed consistently well for any of the performance measures reported by evaluation studies (Table 10).

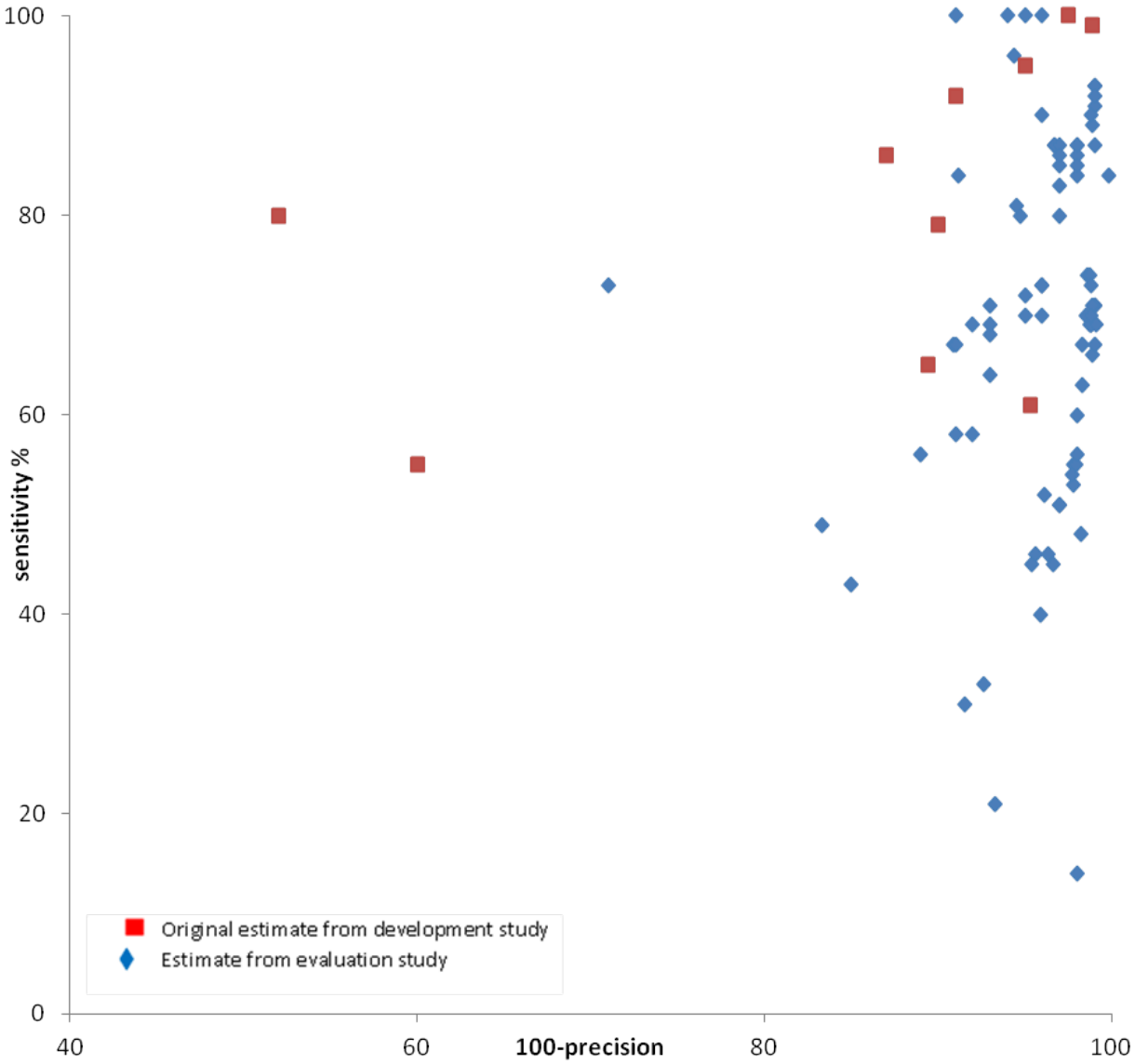
Seven filters had data on both sensitivity and specificity from the original development study and at least one evaluation study (Figure 2). Original estimates showed greater sensitivity and specificity than the estimates from the evaluation studies. The results from the development studies followed a more uniform pattern along a curve, whereas the estimates from the evaluation studies were more heterogenous, especially for specificity. There were two outliers in the evaluation study results: Mitchell's (Mitchell 2005) measure of van der Weijden's (van der Weijden 1997) sensitive filter with very high sensitivity and specificity relative to the other estimates (96% sensitivity; 96% specificity); and Noel-Storr's (Noel-Storr 2011) measure of Haynes 2004 (Haynes 2004) specific filter with very low sensitivity compared to the other estimates (14% sensitivity; 95% specificity). No apparent reason could be found for these anomalous results.

Figure 2. ROC plot of sensitivity and specificity of MEDLINE search filters from development and evaluation studies.



Ten filters had data on both sensitivity and specificity from the original development study and at least one evaluation study (Figure 3). The estimates from both development and evaluation studies showed a wide range in precision and there was substantial variation in sensitivity in the evaluation studies. Precision was generally lower in the evaluation studies, but the pattern was not uniform. There were a number of outliers amongst both the development study and the evaluation study data points. Three outliers had much higher precision than the other estimates. These were: the original performance estimate of the Haynes 1994 specific filter, the original estimate of Deville 2000 strategy 3, and Mitchell's (Mitchell 2005) evaluation of Deville 2000 strategy 4. It was not clear why these precision estimates were high.

Figure 3. ROC plot of sensitivity and precision of MEDLINE search filters from development and evaluation studies.



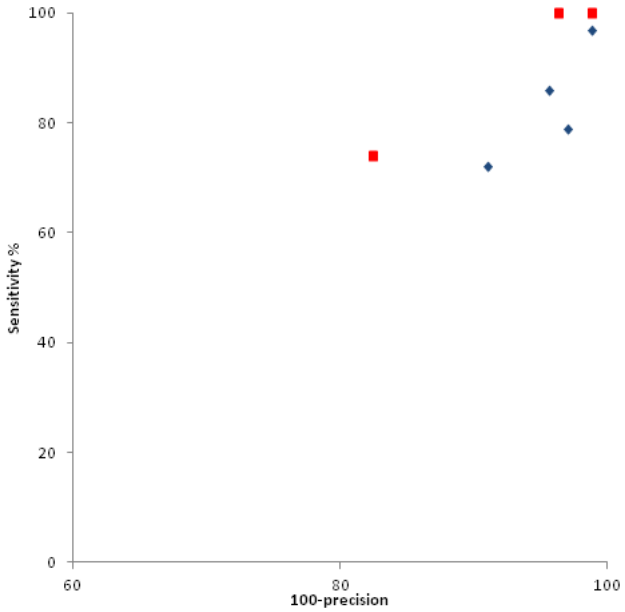
3. Performance of EMBASE filters as reported by development studies

Table 11 shows the 12 filters and their performance data (Bachmann 2003; Wilczynski 2005). Sensitivity ranged from 46% to 100% (median 90%), and precision ranged from 1.2% to 27.7% (median 9%). Half the filters had a sensitivity greater than 90% (median 90.2%), but of these six filters only one had a precision greater than 10 (median 10.4) (Bachmann 2003).

4. Performance of evaluated EMBASE filters

The original studies reported sensitivities ranging from 74% to 100% (median 90%); evaluation studies reporting on the same filters had sensitivities ranging from 72% to 97% (median 86%). The original studies reported precision ranging from 1.2% to 17.6% (median 3.7%); evaluation studies reporting on the same filters had precision ranging from 1.2% to 9% (median 3.7%) (Table 12). One of the evaluated filters did not have an original estimate of performance from the development study (Ovid 2010). Figure 4 shows that in general filters performed better in the original development studies than in the evaluation studies for both sensitivity and precision. None of the filters offered both high sensitivity (> 90%) and high precision (> 10%). The original development studies did not report specificity estimates for the filters that were also tested in evaluation studies, hence a ROC plot of sensitivity and specificity has not been prepared.

Figure 4. ROC plot of sensitivity and precision of EMBASE search filters from development and evaluation studies.



DISCUSSION

Summary of main results

Nineteen studies, reporting 57 MEDLINE filters and 13 EMBASE filters, were eligible for this review. We pre-specified that filters should have a sensitivity > 90% and a precision > 10% to be considered acceptable when searching for studies for systematic reviews of diagnostic test accuracy. We acknowledge that other researchers may set alternative performance levels.

Reports of filter performance were available from studies using a variety of designs, ranging from authors' reports of their filter development process to evaluations of filters carried out by independent researchers using one or more different gold standards. The latter study design should provide best evidence of the performance of filters outside of the original authors' test environment and the consistency of a filter's performance across different sets of records.

Several filters reported performance levels in the development studies which met the pre-specified performance criteria. However, these performance levels typically declined when the filters were validated in the evaluation studies. Thirty MEDLINE filters and four EMBASE filters were tested in an evaluation study against one or more gold standards. In both the evaluation studies that developed their reference set from studies included in several systematic reviews on a broad spectrum of topics, covering a number of publication years, and in those that developed reference sets from heandsearching, no single filter achieved the sensitivity (> 90%) and precision (> 10%) that we pre-specified as 'acceptable'. This means that no filter is suitable for combination with the search terms for the target condition and index tests to create a single search strategy with which to identify studies for systematic reviews of diagnostic test accuracy.

As well as not reaching our pre-specified performance criteria, none of the evaluated filters for use in MEDLINE or EMBASE gave consistent sensitivity and precision measures. This may be caused by translation from one platform to another, or from mistakes made in the transcription of the filters. Another reason may be differences between the indexing and reporting of studies from different scientific fields. For these reasons, the degree of reduction in performance cannot be assessed consistently, making the filters unreliable tools for searching when sensitivity is an important consideration.

Overall completeness and applicability of evidence

The search filters were identified by extensive sensitive searches, checking reference lists of published filters and filter evaluations (Horsley 2011), and by searching a key website which identifies and

collects search filters: the ISSG Search Filter Resource (InterTASC 2011). We are confident that we have identified the vast majority of published search filters, in particular those filters developed using a research method and those tested by independent researchers.

We did not, however, search for unpublished search filters, such as those which might have been developed by people conducting systematic reviews of diagnostic test accuracy studies. There are likely to be many unpublished filters reported in the search strategies of such reviews. These 'unpublished' filters could be identified and evaluated against gold standard sets of relevant records, in the same way that published filters have been evaluated. However, the evidence from the evaluations of the many published filters developed using research methods that we have compiled in this review suggests that unpublished filters may be subject to the same difficulties in achieving the pre-specified performance criteria if those filters consist of variants of the search terms used in the published search filters.

Quality of the evidence

The most reliable filter development studies are likely to be those where the authors used handsearched gold standards and tested their filters against internal validation record sets that are different from the record sets used to develop the filters, and externally validated the filters in a real-world topic. In the one study where this occurred, the MEDLINE filter performance was maintained and had a higher sensitivity (Bachmann 2002).

The nature of the most reliable filter evaluation studies is a matter for debate. Testing filters against a handsearched gold standard set of records would seem to be the most reliable technique because it should yield a range of different DTA study types. However, the disadvantage of handsearching is that researchers are often limited to a small number of journals, which limits the generalisability of the evaluation. Handsearching can be limited by a narrow range of topics and publication years and so impede judgments about the generalisability of the search filters to other topics and time periods. Only two evaluations of MEDLINE filters used handsearched reference sets, which were both topic specific (Deville 2000; Mitchell 2005). In those two studies, some filters maintained their sensitivity as reported in their development papers and others experienced large drops in sensitivity.

Another method of reference set development is to use the studies included in systematic reviews. Whereas handsearching of journals for reference set studies is limited to a small number of journals, using systematic reviews broadens the journal base and the number of publication years covered. However, the primary diagnostic studies in systematic reviews may have been retrieved using a search strategy containing diagnostic terms, which could introduce bias. By including systematic reviews that used a methodological filter to find diagnostic studies, the performance of the evaluated filters in the reference set may be exaggerated. Precision is improved because irrelevant records will be removed but sensitivity may suffer because 'difficult to find' studies may not be retrieved by the filter. This was discussed by Leeflang et al who also used reviews to compile their reference set for the evaluation of 12 filters (Leeflang 2006). Only seven of the reviews in the initial set of 28 reviews used in their study reported search terms. If those seven reviews also used one of the search filters evaluated by Leeflang et al, then the results are likely to be overestimated and the real percentage of missed studies could be even higher than reported. Three other evaluation studies used systematic reviews to compile the reference set, and some of these reviews had included a DTA methodological filter in the original search for eligible studies (Doust 2005; Kastner 2009; Vincent 2003).

How the reference set is used can be a source of bias. If the records used to derive the search terms for the filter are also included in the set of references used in the validation process, this can introduce bias by artificially inflating performance. A discrete set of derivation records and validation records should be used to avoid this. Only two MEDLINE development studies (but neither EMBASE development study) used this approach (Astin 2008; Bachmann 2002).

External validation relates to how generalisable (applicable) the results are to searching for diagnostic studies for different systematic review topics, and only applies to development studies. Four MEDLINE studies carried out external validation of their filters in a real-world setting and were judged to have low concerns about applicability (Bachmann 2002; Deville 2000; Deville 2002; van der Weijden 1997).

The date of the filter may raise concerns. The problem of missed studies is increased in older studies, as shown by Haynes et al whose filter tested in the 1986 reference set did not perform as well as it did in the 1991 reference set. This may be a feature of the reporting of DTA studies. The STARD statement, which was published in 2003, aimed to improve the standard of reporting of DTA studies (Bossuyt 2003). STARD's first recommendation is that authors should identify their publication as a study of diagnostic accuracy. If authors and editors support STARD, this alone will enhance the efficient retrievability of DTA studies.

There are concerns that the same filter may not have been implemented uniformly across evaluation studies and that this may hamper an evaluation of the consistency of filter performance. Some researchers have translated filters across searching platforms, for example from Ovid MEDLINE to PubMed. The translation process may influence the performance of the filters, although the likely effect of this is unclear. Translations may change the number of missed studies and may impact sensitivity and precision. PubMed, in particular, carries out automatic mapping of search terms and this factor needs to be taken into account when translating from PubMed to other interfaces and when translating a strategy to make it suitable for use in PubMed. An example is the different adaptations made by the Haynes team in translating the original Haynes 1994 sensitive search filter developed in Ovid into the PubMed Clinical Queries sensitive filter, and Falck-Ytter's adaptation of the same filter for use in PubMed. Sensitivities reported by the evaluation studies varied between each of the three filters, which may be due to differences in translation. Furthermore, some evaluators report strategies with mistakes; the mistakes might have been made in the conduct of the strategies or might have been introduced at the reporting stage. This uncertainty leads to doubts about the performance data reported for some of the filters, and we were unable to make any judgement about whether the original filters were applied correctly in the evaluation studies.

Potential biases in the review process

It can be difficult to identify the filters reported in evaluation studies because the filters can be named differently and the filter used is

not always listed in the paper or appendix (that is only a reference may be provided). In those circumstances, it is unclear whether the strategy was used accurately or whether it was adapted. In some cases the original source of a filter has disappeared because of changes to websites. The Shipley Miner and University of Rochester filters evaluated by Ritchie, Whiting and Vincent are no longer available online and we have to rely on the evaluators for a listing of the strategies, rather than being able to visit the original website. This means that our review may have erroneously assigned some performance data to a named filter or to a filter which is a variant of a published filter.

Agreements and disagreements with other studies or reviews

Many of the search filters included in this systematic review have been extensively evaluated in other studies with different but relevant gold standards. This systematic review of evaluation studies draws the same conclusions as the most comprehensive evaluation study by Whiting and colleagues, which concluded that filtered searches miss additional studies compared with searches based on index test and target condition alone (Whiting 2010). None of the filters evaluated by Whiting provided reductions in Number Needed to Read for acceptable sensitivity and should not be used to identify studies for inclusion in systematic reviews (Whiting 2010). A key strength of the Whiting study is the size and homogeneity in the creation of the reference set; the team used seven systematic reviews published on a broad range of topics that had been conducted by the authors using extensive, rigorous and, for the first time, reproducible search methods without the inclusion of a methodological search filter. The inclusion criteria for each review produced sufficient data to allow cross-tabulation of results comparing index tests with a reference standard and meant that only true test accuracy studies were included.

AUTHORS' CONCLUSIONS

Implication for methodological research

The information retrieval environment is not static and better reporting of DTA studies as advocated by STARD, additional indexing terms (as recently introduced by EMBASE) and more consistent indexing of diagnostic studies could help to make published methodological filters more sensitive or create the opportunity for the development of new filters.

Search filters which make more use of proximity operators and careful exclusion may also yield improvements in performance in traditional database interfaces reliant on Boolean searching. Beyond Boolean approaches, developments in information retrieval such as semantic textual analysis may lead to filtering programs or record matching rules which can better identify diagnostic test accuracy studies from batches of records retrieved by sensitive searches. The increasing availability of full text journals may also improve the retrieval of DTA studies as there will be the whole paper to search and DTA performance measures may be more consistently identified.

In the absence of current suitable search filters, the impact of different search approaches could be investigated. The effectiveness of multi-strand searching is unexplored. In addition, the yield of restricted searching on the results of the systematic review could be explored. A combination of search approaches where the results of strategies using filters are augmented with more extensive reference checking and citation searching could also be investigated as an alternative approach to identifying as many relevant DTA studies as possible.

ACKNOWLEDGEMENTS

We thank Marit Johansen for her help in designing the search strategies.

REFERENCES

References to studies included in this review

Astin 2008 *{published data only}*

Astin MP, Brazzelli MG, Fraser CM, Counsell CE, Needham G, Grimshaw JM. Developing a sensitive search strategy in MEDLINE to retrieve studies on assessment of the diagnostic performance of imaging techniques. *Radiology* 2008;**247**(2):365-73. [MEDLINE: 18372447]

Bachmann 2002 *{published data only}*

Bachmann LM, Coray R, Estermann P, Ter Riet G. Identifying diagnostic studies in MEDLINE: reducing the number needed to read. *Journal of the American Medical Informatics Association* 2002;**9**(6):653-8. [MEDLINE: 12386115]

Bachmann 2003 *{published data only}*

Bachmann LM. Identifying diagnostic accuracy studies in EMBASE. *Journal of the Medical Library Association* 2003;**91**(3):341-6. [MEDLINE: 12883560]

Berg 2005 *{published data only}*

Berg A, Fleischer S, Behrens J. Development of two search strategies for literature in MEDLINE-PubMed: nursing diagnoses in the context of evidence-based nursing. *International Journal of Nursing Terminologies and Classifications* 2005;**16**(2):26-32. [MEDLINE: 16045550]

Deville 2000 *{published data only}*

Deville WL, Bezemer PD, Bouter LM. Publications on diagnostic test evaluation in family medicine journals: an optimal search strategy. *Journal of Clinical Epidemiology* 2000;**53**(1):65-9. [MEDLINE: 10693905]

Deville 2002 *{published data only}*

Deville WL, Bossuyt PM, de Vet HC, Bezemer PD, Bouter LM, Assendelft WJ. Systematic reviews in practice. X. Searching, selecting and the methodological assessment of diagnostic evaluation research [De praktijk van systematische reviews. X. Zoeken, selecteren en methodologisch beoordelen van diagnostisch evaluatieonderzoek]. *Nederlands Tijdschrift voor Geneeskunde* 2002;**146**(48):2281-4. [MEDLINE: 12497754]

Doust 2005 *{published data only}*

Doust JA, Pietrzak E, Sanders S, Glasziou PP. Identifying studies for systematic reviews of diagnostic tests was difficult due to the poor sensitivity and precision of methodologic filters and the lack of information in the abstract. *Journal of Clinical Epidemiology* 2005;**58**(5):444-9. [MEDLINE: 15845330]

Haynes 1994 *{published data only}*

* Haynes RB, Wilczynski, McKibbon KA, Walker CJ, Sinclair JC. Developing optimal search strategies for detecting clinically sound studies in MEDLINE. *Journal of the American Medical Informatics Association* 1994;**1**(6):447-58. [MEDLINE: 7850570]

Wilczynski NL, Walker CJ, McKibbon KA, Haynes RB. Assessment of methodologic search filters in MEDLINE. *Proceedings of the Annual Symposium on Computer Applications in Medical Care* 1993:601-5.

Wilczynski NL, Walker CJ, McKibbon KA, Haynes RB. Quantitative comparison of pre-explosions and subheadings with methodologic search terms in MEDLINE. *Proceedings of the Annual Symposium on Computer Applications in Medical Care* 1994:905-9.

Haynes 2004 *{published data only}*

Haynes RB, Wilczynski NL. Optimal search strategies for retrieving scientifically strong studies of diagnosis from Medline: analytical survey. *BMJ* 2004;**328**(7447):1040. [MEDLINE: 15073027]

Kassai 2006 *{published data only}*

Kassai B, Sonie S, Shah NR, Boissel JP. Literature search parameters marginally improved the pooled estimate accuracy for ultrasound in detecting deep venous thrombosis. *Journal of Clinical Epidemiology* 2006;**59**(7):710-4. [MEDLINE: 16765274]

Kastner 2009 *{published data only}*

Kastner M, Wilczynski NL, McKibbon AK, Garg AX, Haynes RB. Diagnostic test systematic reviews: bibliographic search filters ("Clinical Queries") for diagnostic accuracy studies perform well. *Journal of Clinical Epidemiology* 2009;**62**(9):974-81. [MEDLINE: 19230607]

Leeflang 2006 *{published data only}*

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. *Journal of Clinical Epidemiology* 2006;**59**(3):234-40. [MEDLINE: 16488353]

Mitchell 2005 *{published data only}*

Mitchell RL, Rinaldi F, Craig JC. Performance of published search strategies for studies of diagnostic test accuracy (SDTAs) in MEDLINE and EMBASE. XIII Cochrane Colloquium; 22-26 Melbourne, Australia. 2005.

Noel-Storr 2011 *{published data only}*

Noel-Storr A. The development of a methodological filter for studies of diagnostic accuracy in dementia. IXX Cochrane Colloquium, 19-22 October Madrid, Spain. 2011.

Ritchie 2007 *{published data only}*

Ritchie G, Glanville J, Lefebvre C. Do published search filters to identify diagnostic test accuracy studies perform adequately?. *Health Information and Libraries Journal* 2007;**24**(3):188-92. [MEDLINE: 17714173]

van der Weijden 1997 *{published data only}*

van der Weijden T, IJzermans CJ, Dinant GJ, van Duijn NP, de Vet R, Buntinx F. Identifying relevant diagnostic studies in MEDLINE. The diagnostic value of the erythrocyte sedimentation rate (ESR) and dipstick as an example. *Family Practice* 1997;**14**(3):204-8. [MEDLINE: 9201493]

Vincent 2003 *{published data only}*

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 Information and Libraries Journal* 2003;**20**(3):150-9. [MEDLINE: 12919278]

Whiting 2010 *{published data only}*

Whiting P, Westwood M, Beynon R, Burke M, Sterne JA, Glanville J. Inclusion of methodological filters in searches for diagnostic test accuracy studies misses relevant studies. *Journal of Clinical Epidemiology* 2010;**64**(6):602-7. [MEDLINE: 21075596]

Wilczynski 2005 *{published data only}*

Wilczynski NL, Haynes RB, Hedges Team. EMBASE search strategies for identifying methodologically sound diagnostic studies for use by clinicians and researchers. *BMC Medicine* 2005;**3**:7. [MEDLINE: 15796772]

Additional references

Bossuyt 2003

Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. *Clinical Chemistry* 2003;**49**(1):7-18. [MEDLINE: 12507954]

CASP 2002

Critical Appraisal Skills Programme. Search Filters. http://www.phru.nhs.uk/casp/search_filters.htm (No longer available) 2006.

Deeks 2010

Deeks JJ, Bossuyt PM, Gatsonis C (editors). Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 1.0. The Cochrane Collaboration, 2010. Available from http://srdta.cochrane.org (accessed 25 April 2013).

DeVet 2008

de Vet HCW, Eisinga A, Riphagen II, Aertgeerts B, Pewsner D, Mitchell R. Chapter 7: Searching for studies. In: Deeks JJ, Bossuyt PM, Gatsonis C (editors). Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 0.4 [updated September 2008]. The Cochrane Collaboration, 2008. Available from http://srdta.cochrane.org (accessed 25 April 2013).

Deville 2002a

Deville WL, Buntinx F, Bouter LM, Montori VM, De Vet HC, van der Windt DA. Conducting systematic reviews of diagnostic studies: didactic guidelines. *BMC Medical Research Methodology* 2002;**2**:9.

Falck-Ytter 2004

Falck-Ytter YT, Motschall E. New search filter for diagnostic studies: Ovid and PubMed versions not the same [2004]. available at http://www.bmj.com/content/328/7447/1040?tab=responses (accessed 25 April 2013).

Fielding 2002

Fielding AM, Powell A. Using Medline to achieve an evidence-based approach to diagnostic clinical biochemistry. *Annals of Clinical Biochemistry* 2002;**39**(Pt 4):345-50. [MEDLINE: 12117438]

Glanville 2008

Glanville J, Bayliss S, Booth A, Dundar Y, Fleeman ND, Foster L, et al. on behalf of the InterTASC Information Specialists' Subgroup. So many filters, so little time: the development of a search filter appraisal checklist. *Journal of the Medical Library Association* 2008;**96**(4):356-61. [MEDLINE: 18974813]

Haynes 2005

Haynes RB, McKibbon KA, Wilczynski NL, Walter SD, Werre SR, Hedges Team. Optimal search strategies for retrieving scientifically strong studies of treatment from Medline: analytical survey. *BMJ* 2005;**330**(7501):1179-84. [MEDLINE: 15894554]

Haynes 2005a

Haynes RB, Kastner M, Wilczynski NL, Hedges Team. Developing optimal search strategies for detecting clinically sound and relevant causation studies in EMBASE. *BMC Medical Information and Decision Making* 2005;**5**:8. [MEDLINE: 15784134]

Horsley 2011

Horsley T, Dingwall O, Sampson M. Checking reference lists to find additional studies for systematic reviews. *Cochrane Database of Systematic Reviews* 2011, Issue 8. [DOI: 10.1002/14651858.MR000026.pub2]

InterTASC 2011

InterTASC Information Specialists' Sub-Group (ISSG). The InterTASC Information Specialists' Sub-Group Search Filter Resource: diagnostic studies. Available at http://www.york.ac.uk/inst/crd/intertasc/diag.htm. York: Centre for Reviews and Dissemination, (accessed 25th April 2013).

Lefebvre 2011

Lefebvre C, Manheimer E, Glanville J. Chapter 6: Searching for studies. In: Higgins JPT, Green S (editors). Cochrane Handbook for Systematic Reviews of Interventions. Version 5.1.0 [updated March 2011]. The Cochrane Collaboration, 2011. Available from http://www.cochrane-handbook.org (accessed 25 April 2013).

NLM 2005

US National Library of Medicine. Clinical Queries using Research Methodology Filters [updated Jan 2005]. Available from http://www.ncbi.nlm.nih.gov/books/NBK3827/#pubmedhelp.Clinical_Queries_Filters (accessed 25 April 2013).

North Thames 2002

North Thames. Diagnostic procedures. http://www.londonlinks.ac.uk/evidence_strategies/ovid_filters.htm#diagnostic. (No longer available).

Ovid 2010

Wolfer Kluer Health. Clinical queries in Ovid. available at: http://ovidsupport.custhelp.com/cgi-bin/ovidsupport.cfg/php/enduser/std_adp.php?p_faqid=1599&;p_created=1087487498&p_sid=B86UCj8j&p_accessibility=0&p_redirect=&p_lva=&p_sp=cF9zcmNoPTEmcF9zb3J0X2J5PSZwX2dyaWRzb3J0PSZwX3Jvd19jbnQ9I li=&p_topview=1 First published 2004; updated 2010.

OvidSP 2013

Wolters Kluwer Health. MEDLINE® 2013 Database Guide. available at http://ovidsp.tx.ovid.com/sp-3.8.1a/ovidweb.cgi?&S=BLIMFPMFDFDDHIFFNCOKFBFBJDCBAA00&Database+Field+Guide=37 [2012] (accessed 25 April 2013).

OvidSP 2013a

Wolters Kluwer Health. Embase: Excerpta Medica Database Guide. http://ovidsp.tx.ovid.com/sp-3.8.1a/ovidweb.cgi?&S=BLIMFPMFDFDDHIFFNCOKFBFBJDCBAA00&Database+Field+Guide=10 [2012] (accessed 25 April 2013).

Shipley Miner 2002

Shipley MC. Evidence based filters for Ovid MEDLINE. http://www.urmc.rochester.edu/hslt/miner/digital_library/tip_sheets/OVID_eb_filters.pdf. Rochester: Edward G Miner Library, University of Rochester.

University of Rochester 2002

Miner Library Reference Librarians. Evidence based filters for Ovid MEDLINE. Miner Library, University of Rochester 2002.

Whiting 2008

Whiting P, Westwood M, Burke M, Sterne J, Harbord R, Glanville J. Can diagnostic filters offer similar sensitivity and a reduced number needed to read compared to searches based on index test and target condition? [abstract]. Methods for Evaluating Medical Tests. Symposium. 2008 Jul 24-25.

Whiting 2011

Whiting P, Westwood M, Beynon R, Burke M, Sterne JA, Glanville J. Inclusion of methodological filters in searches for diagnostic accuracy studies misses relevant studies. *Journal of Clinical Epidemiology* 2011;**64**(6):602-7. [MEDLINE: 21075596]

Wilczynski 1995

Wilczynski NL, Walker CJ, McKibbon KA, Haynes RB. Reasons for the loss of sensitivity and specificity of methodologic MeSH terms and textwords in MEDLINE. *Proceedings - the Annual Symposium on Computer Applications in Medical Care* 1995:436-40. [MEDLINE: 8563319]

Wilczynski 2003

Wilczynski NL, Haynes RB, Hedges Team. Developing optimal search strategies for detecting clinically sound causation studies in MEDLINE. *AMIA - Annual Symposium Proceedings/AMIA Symposium* 2003:719-23. [MEDLINE: 14728267]

Wilczynski 2004

Wilczynski NL, Haynes RB, Hedges Team. Developing optimal search strategies for detecting clinically sound prognostic studies in MEDLINE: an analytic survey. *BMC Medicine* 2004;**2**(1):23. [MEDLINE: 15189561]

Wilczynski 2005a

Wilczynski NL, Haynes RB. Optimal search strategies for detecting clinically sound prognostic studies in EMBASE: an analytic survey. *Journal of the American Medical Informatics Association* 2005;**12**(4):481-5. [MEDLINE: 15802476]

Wilczynski 2007

Wilczynski NL, Haynes RB. Indexing of diagnosis accuracy studies in MEDLINE and EMBASE. *AMIA - Annual Symposium Proceedings/AMIA Symposium* 2007:801-5. [MEDLINE: 18693947]

* Indicates the major publication for the study

CHARACTERISTICS OF STUDIES

Characteristics of included studies *[ordered by study ID]*

Astin 2008

Methods	**Method of identification of reference set records -** Handsearching
	Method of deriving filter terms - Analysis of reference set
Data	**Reference set years -** Development set 1985 Clin Radiol, 1988 Am J Neuroradiol; validation set 2000
	Number of gold standard records - 333 in development set; 186 in validation set
	Number of non-gold standard records - 2222 in development set; 1070 in validation set
Comparisons	**Reference set also contained non-gold standard records -**Yes
	Description of non-gold standard records if used in reference set - Not reported
Outcomes	**Number of filters developed -** 1
Notes	MEDLINE development study
Risk of bias	

Astin 2008 *(Continued)*

Item	Authors' judgement	Description
If relevant, systematic review did not use DTA strategy?	Unclear	Not relevant
Generic gold-standard records?	No	Filter developed to retrieve radiology DTA studies. High concerns about applicability
Independent internal validation?	Yes	Discrete set of references in a derivation set from six handsearched journals and a validation set from six different handsearched journals
Externally validated?	No	High concerns about applicability

Bachmann 2002

Methods	**Method of identification of reference set records -** Handsearching **Method of deriving filter terms -** Analysis of reference set
Data	**Reference set years -** 1989, 1994 and 1999 **Number of gold standard records -** 83 in 1989 test set; 53 in 1994 validation set; 61 in 1999 validation set **Number of non-gold standard records -** 1646 in 1989 test set; 1744 in 1994 validation set; 7875 in 1999 validation set
Comparisons	**Reference set also contained non-gold standard records -** Yes **Description of non-gold standard records if used in reference set -** Not reported
Outcomes	**Number of filters developed -** 1
Notes	MEDLINE development study

Risk of bias

Item	Authors' judgement	Description
If relevant, systematic review did not use DTA strategy?	Unclear	Not relevant, systematic review not used
Generic gold-standard records?	Yes	Low concerns about applicability
Independent internal validation?	Yes	Terms derived from 1989 reference set; filter validated in 1994 validation set
Externally validated?	Yes	References from search of different journals and year to the derivation and internal validation set. Low concerns about applicability

Bachmann 2003

Methods	**Method of identification of reference set records -** Handsearching **Method of deriving filter terms -** Analysis of reference set
Data	**Reference set years -** 1999 **Number of gold standard records -** 61 **Number of non-gold standard records -** 6082
Comparisons	**Reference set also contained non-gold standard records -** Yes **Description of non-gold standard records if used in reference set -** All records retrieved by search that were not classified as gold standard studies
Outcomes	**Number of filters developed -** 8
Notes	EMBASE development study

Risk of bias

Item	Authors' judgement	Description
If relevant, systematic review did not use DTA strategy?	Unclear	Not relevant
Generic gold-standard records?	Yes	Low concerns about applicability
Independent internal validation?	No	Terms for filters derived through word frequency analysis of the same references as the validation set
Externally validated?	No	

Berg 2005

Methods	**Method of identification of reference set records -** Handsearching **Method of deriving filter terms -** Analysis of reference set and adaption of existing filter
Data	**Reference set years -** Not reported **Number of gold standard records -** Not reported **Number of non-gold standard records -** 238
Comparisons	**Reference set also contained non-gold standard records -** Yes **Description of non-gold standard records if used in reference set -** Not reported
Outcomes	**Number of filters developed -** 1
Notes	MEDLINE development study

Risk of bias

Berg 2005 *(Continued)*

Item	Authors' judgement	Description
If relevant, systematic review did not use DTA strategy?	Unclear	Not relevant
Generic gold-standard records?	No	Cancer-related fatigue topic specific. High concerns about applicability
Independent internal validation?	No	Used the indexing of included citations from gold standard references to derive terms, these references also included in validation
Externally validated?	No	High concerns about applicability

Deville 2000

Methods	**Method of identification of reference set records -** Handsearching **Method of deriving filter terms -** Analysis of reference set
Data	**Reference set years -** 1992-1995 **Number of gold standard records -** 75; 33 in meniscal lesions set **Number of non-gold standard records -** 2392; meniscal lesions set not reported
Comparisons	**Reference set also contained non-gold standard records -** Yes **Description of non-gold standard records if used in reference set -** False positive papers selected by a previously published search strategy, exclusion of some publication types (e.g. reviews and meta-analyses)
Outcomes	**Number of filters developed -** 4 **Number of filters evaluated -** 1
Notes	MEDLINE development and evaluation study

Risk of bias

Item	Authors' judgement	Description
If relevant, systematic review did not use DTA strategy?	Unclear	Not relevant
Generic gold-standard records?	No	Family medicine reference set; physical diagnostic tests for meniscal lesion validation set. High concerns about applicability
Independent internal validation?	No	
Externally validated?	Yes	

Deville 2002

Methods	**Method of identification of reference set records -** DTA systematic reviews
	Method of deriving filter terms - Adaption of existing filter
Data	**Reference set years -** Not reported
	Number of gold standard records - Not reported
	Number of non-gold standard records - Not reported
Comparisons	**Reference set also contained non-gold standard records -** Not reported
	Description of non-gold standard records if used in reference set - Not reported
Outcomes	**Number of filters developed -** 1
Notes	MEDLINE development study

Risk of bias

Item	Authors' judgement	Description
If relevant, systematic review did not use DTA strategy?	Unclear	The reference cited by the study to the systematic review which was used is unavailable. A meta-analysis published by the same author on the topic did use a search strategy containing diagnostic terms
Generic gold-standard records?	No	Studies from a systematic review of diagnostic tests for knee lesions and a systematic review of a urine dipstick test comprised the reference set. High concerns about applicability
Externally validated?	Yes	Real-world validation sets based on two systematic reviews. Low concerns about applicability

Doust 2005

Methods	**Method of identification of reference set records -** DTA systematic reviews conducted by authors
Data	**Reference set years -** Tympanometry 1966-2001; natriuretic peptides 1994-2002
	Number of gold standard records - Tympanometry n=33; natriuretic peptides n=20
	Number of non-gold standard records - 0
Comparisons	**Reference set also contained non-gold standard records -** No
	Description of non-gold standard records if used in reference set - Not applicable
Outcomes	**Number of filters evaluated -** 5
Notes	MEDLINE evaluation study

Risk of bias

Item	Authors' judgement	Description

Doust 2005 *(Continued)*

If relevant, systematic review did not use DTA strategy?	No	The authors conducted two systematic reviews whose studies comprised the reference set. The clinical queries filter for diagnostic studies available in PubMed was used.
Generic gold-standard records?	No	Studies from a systematic review of tympanometry for the diagnosis of otitis media with effusion in children and a systematic review of natriuretic peptides comprised the reference standard. High concerns about applicability
Independent internal validation?	Unclear	Not relevant
Externally validated?	Unclear	Not relevant

Haynes 1994

Methods	**Method of identification of reference set records -** Handsearching for primary studies **Method of deriving filter terms -** Expert knowledge and analysis of reference set
Data	**Reference set years -** 1986 and 1991 **Number of gold standard records -** 92 in 1986 set; 111 in 1991 set **Number of non-gold standard records -** 426 in 1986 set; 301 in 1991 set
Comparisons	**Reference set also contained non-gold standard records -** Yes **Description of non-gold standard records if used in reference set -** Not reported
Outcomes	**Number of filters developed -** 12
Notes	MEDLINE development study. All papers listed under Haynes 1994 used for data extraction

Risk of bias

Item	Authors' judgement	Description
If relevant, systematic review did not use DTA strategy?	Unclear	Not relevant
Generic gold-standard records?	Yes	Low concerns about applicability
Independent internal validation?	No	Terms were collected through expert knowledge but their combination into a strategy was not independent of the references used for validation. The reference standard was used to eliminate terms with <10% sensitivity or combination with <40% sensitivity or <70% specificity
Externally validated?	No	High concerns about applicability

Haynes 2004

Methods	**Method of identification of reference set records -** Handsearching for primary studies

Haynes 2004 *(Continued)*

	Method of deriving filter terms - Expert knowledge and analysis of reference set
Data	**Reference set years -** 2000 **Number of gold standard records -** 147 **Number of non-gold standard records -** 48,881
Comparisons	**Reference set also contained non-gold standard records -** Yes **Description of non-gold standard records if used in reference set -** Not reported
Outcomes	**Number of filters developed -** 11
Notes	MEDLINE development study

Risk of bias

Item	Authors' judgement	Description
If relevant, systematic review did not use DTA strategy?	Unclear	Not relevant
Generic gold-standard records?	Yes	Low concerns about applicability
Independent internal validation?	No	Individual search terms with a sensitivity >25% and a specificity >75% (when tested in the reference set) were incorporated into the development of search strategies
Externally validated?	No	High concerns about applicability

Kassai 2006

Methods	**Method of identification of reference set records -** Primary studies identified through Internet search
Data	**Reference set years -** 1966-2002 **Number of gold standard records -** 237 **Number of non-gold standard records -** 1236
Comparisons	**Reference set also contained non-gold standard records -** Yes **Description of non-gold standard records if used in reference set -** All studies retrieved by search not classified as gold standard records
Outcomes	**Number of filters evaluated -** 3
Notes	MEDLINE evaluation study

Risk of bias

Item	Authors' judgement	Description

Kassai 2006 *(Continued)*

If relevant, systematic review did not use DTA strategy?	Unclear	Not relevant
Generic gold-standard records?	No	Venous thrombosis and ultrasonography topic specific. High concerns about applicability
Independent internal validation?	Unclear	Not relevant
Externally validated?	Unclear	Not relevant

Kastner 2009

Methods	**Method of identification of reference set records -** Internet search for DTA systematic reviews
Data	**Reference set years -** 2006 **Number of gold standard records -** 441 **Number of non-gold standard records -** 0
Comparisons	**Reference set also contained non-gold standard records -** No **Description of non-gold standard records if used in reference set -** Not applicable
Outcomes	**Number of filters evaluated -** 1
Notes	MEDLINE evaluation study

Risk of bias

Item	Authors' judgement	Description
If relevant, systematic review did not use DTA strategy?	No	Five of the twelve systematic reviews which provided studies for the reference set, used search strategies containing DTA search terms to find primary studies
Generic gold-standard records?	Yes	Low concerns about applicability
Independent internal validation?	Unclear	Not relevant
Externally validated?	Unclear	Not relevant

Leeflang 2006

Methods	**Method of identification of reference set records -** Internet search for DTA systematic reviews
Data	**Reference set years -** 1999-2002 **Number of gold standard records -** 820

Leeflang 2006 *(Continued)*

	Number of non-gold standard records - 0
Comparisons	**Reference set also contained non-gold standard records -** No
	Description of non-gold standard records if used in reference set - Not applicable
Outcomes	**Number of filters evaluated -** 12
Notes	MEDLINE evaluation study

Risk of bias

Item	Authors' judgement	Description
If relevant, systematic review did not use DTA strategy?	Unclear	Of the 27 systematic reviews whose studies were used to comprise the reference set, seven did not describe their search strategy. It is unclear, therefore, whether diagnostic terms would have been applied in the search.
Generic gold-standard records?	Yes	Low concerns about applicability
Independent internal validation?	Unclear	Not relevant
Externally validated?	Unclear	Not relevant

Mitchell 2005

Methods	**Method of identification of reference set records -** Handsearching for primary studies
Data	**Reference set years -** 1991-1992; 2002-2003
	Number of gold standard records - 99
	Number of non-gold standard records - 4409
Comparisons	**Reference set also contained non-gold standard records -** Yes
	Description of non-gold standard records if used in reference set - Not reported
Outcomes	**Number of filters evaluated -** 6 MEDLINE filters and 6 EMBASE filters
Notes	MEDLINE and EMBASE evaluation study

Risk of bias

Item	Authors' judgement	Description
If relevant, systematic review did not use DTA strategy?	Unclear	Not relevant
Generic gold-standard records?	No	Kidney disease topic specific. High concerns about applicability

Mitchell 2005 *(Continued)*

Independent internal validation?	Unclear	Not relevant
Externally validated?	Unclear	Not relevant

Noel-Storr 2011

Methods	**Method of identification of reference set records -** DTA systematic reviews conducted by the authors **Method of deriving filter terms -** Analysis of reference set (authors ran published search filters in MEDLINE combined with a subject search, locating 10 papers that all filters missed and choosing a term from their title/abstract or keywords of each)
Data	**Reference set years -** 2000-2001 **Number of gold standard records -** 128 in September 2010 set with additional 16 found in update search. Therefore, 144 in August 2011 **Number of non-gold standard records -** 17,266 in September 2010 set; with additional 1654 found in update search; so 18,920 in August 2011
Comparisons	**Reference set also contained non-gold standard records -** Yes **Description of non-gold standard records if used in reference set -** All studies retrieved by search not classified as gold standard records
Outcomes	**Number of filters developed -** 1
Notes	MEDLINE development and evaluation study

Risk of bias

Item	Authors' judgement	Description
If relevant, systematic review did not use DTA strategy?	Yes	
Generic gold-standard records?	No	Studies included in a systematic review of biomarkers for diagnosing mild cognitive impairment comprised reference set; filter designed to retrieve longitudinal DTA studies. High concerns about applicability
Independent internal validation?	No	The reference was not totally independent of the set used to derive terms, it consisted of 144 gold standard records and 18,920 non-gold standard records, but the 10 studies used to derive terms for the new filter were included in the reference set during validation
Externally validated?	No	High concerns about applicability

Ritchie 2007

Methods	**Method of identification of reference set records -** Primary studies identified through Internet search
Data	**Reference set years -** 1966-2003

Ritchie 2007 *(Continued)*

	Number of gold standard records - 160
	Number of non-gold standard records - 27,804
Comparisons	**Reference set also contained non-gold standard records -** Yes
	Description of non-gold standard records if used in reference set - All studies retrieved by search not classified as gold standard records
Outcomes	**Number of filters evaluated -** 22
Notes	MEDLINE evaluation study

Risk of bias

Item	Authors' judgement	Description
If relevant, systematic review did not use DTA strategy?	Unclear	Not relevant
Generic gold-standard records?	No	Childhood urinary tract infection diagnosis topic specific. High concerns about applicability
Independent internal validation?	Unclear	Not relevant
Externally validated?	Unclear	Not relevant

van der Weijden 1997

Methods	**Method of identification of reference set records -** Personal literature database
	Method of deriving filter terms - Checking key publications for terms and language used
Data	**Reference set years -** 1985-1994
	Number of gold standard records - 221
	Number of non-gold standard records - 0
Comparisons	**Reference set also contained non-gold standard records -** No
	Description of non-gold standard records if used in reference set - Not applicable
Outcomes	**Number of filters developed -** 3
Notes	MEDLINE development study

Risk of bias

Item	Authors' judgement	Description
If relevant, systematic review did not use DTA strategy?	Unclear	Not relevant

van der Weijden 1997 *(Continued)*

Generic gold-standard records?	No	Erythrocyte sedimentation as a diagnostic test topic specific. High concerns about applicability
Independent internal validation?	No	Filters composed of terms that were derived from checking the key publications for terms and language used, not judged to be internally validated as only real-world external validation carried out
Externally validated?	Yes	Two systematic reviews on ESR and dipstick testing provided references for validation testing. Low concerns about applicability

Vincent 2003

Methods	**Method of identification of reference set records -** DTA systematic reviews **Method of deriving filter terms -** Adaption of existing filter and analysis of reference set
Data	**Reference set years -** 1969-2000 **Number of gold standard records -** 126 **Number of non-gold standard records -** 0
Comparisons	**Reference set also contained non-gold standard records -** No **Description of non-gold standard records if used in reference set -** Not applicable
Outcomes	**Number of filters developed -** 3 **Number of filters evaluated -** 5
Notes	MEDLINE development and evaluation study

Risk of bias

Item	Authors' judgement	Description
If relevant, systematic review did not use DTA strategy?	No	At least one of the 16 systematic reviews used to provide studies for the reference set, used diagnostic search terms in the search strategy. Many of the systematic reviews did not provide a full search strategy and, therefore, it is unclear whether they would have used a diagnostic filter or not.
Generic gold-standard records?	No	Deep vein thrombosis diagnosis topic specific. High concerns about applicability
Independent internal validation?	No	Published filters were adapted by adding and removing terms based on the results of searches of the reference set
Externally validated?	No	High concerns about applicability

Whiting 2010

Methods	**Method of identification of reference set records -** Systematic reviews conducted by the authors
Data	**Reference set years -** Not reported

Whiting 2010 *(Continued)*

	Number of gold standard records - 506
	Number of non-gold standard records - 25,880 (number obtained from authors)
Comparisons	**Reference set also contained non-gold standard records -** Yes
	Description of non-gold standard records if used in reference set - Not reported
Outcomes	**Number of filters evaluated -** 22
Notes	MEDLINE evaluation study

Risk of bias

Item	Authors' judgement	Description
If relevant, systematic review did not use DTA strategy?	Yes	The authors conducted the systematic reviews and state that their search strategies did not contain any diagnostic terms
Generic gold-standard records?	Yes	DTA studies from seven systematic reviews which covered a range of different types of diagnostic test and condition. Low concerns about applicability
Independent internal validation?	Unclear	Not relevant
Externally validated?	Unclear	Not relevant

Wilczynski 2005

Methods	**Method of identification of reference set records -** Handsearching for primary studies
	Method of deriving filter terms - Analysis of reference set and expert knowledge
Data	**Reference set years -** 2000
	Number of gold standard records - 97
	Number of non-gold standard records - 27,672
Comparisons	**Reference set also contained non-gold standard records -** Yes
	Description of non-gold standard records if used in reference set - Not reported
Outcomes	**Number of filters developed -** 4
Notes	EMBASE development study

Risk of bias

Item	Authors' judgement	Description
If relevant, systematic review did not use DTA strategy?	Unclear	Not relevant

Wilczynski 2005 *(Continued)*

Generic gold-standard records?	Yes	Low concern about applicability
Independent internal validation?	No	
Externally validated?	No	High concerns about applicability

ADDITIONAL TABLES

Table 1. Summary of study designs of MEDLINE filter development studies

	Author (year)									
	Astin 2008	Berg 2005	van der Weijden 1997	Deville 2002	Deville 2000	Haynes 2004	Haynes 1994	Bachmann 2002	Vincent 2003	Noel-Storr 2011
Method of identification of reference set records (one from list below selected for each study)										
Hand-searching for primary studies	✓	✓	-	-	✓	✓	✓	✓	-	-
DTA systematic reviews	-	-	-	✓	-	-	-	-	✓	✓
Personal literature database	-	-	✓	-	-	-	-	-	-	-
If systematic reviews used in reference set development, did they include DTA search terms in search strategy?										
	-	-	-	Unclear	-	-	-	-	✓	X
Reference set also contained non-gold standard records										
	✓	✓	X	NR	✓	✓	✓	✓	X	✓
Description of non-gold standard records if used in reference set	NR	-	-	-	-	NR	NR	NR	-	-
All studies retrieved by search not classified as gold standard records		✓	-	-	-	-	-	-	-	✓
False positive papers selected by a previously published search strategy, exclusion of some publication types e.g. reviews and meta-analyses.		-	-	-	✓	-	-	-	-	-
Generic gold standard records i.e. not topic specific										
	X	X	X	X	X	✓	✓	✓	X	X

Table 1. Summary of study designs of MEDLINE filter development studies *(Continued)*

Method of deriving filter terms (a combination of methods could be used)										
Analysis of reference set	✓	✓	-	-	✓	✓	✓	✓	✓	✓*
Expert knowledge	-	-	-	-	-	✓	✓	-	-	-
Adaption of existing filter	-	✓	-	✓	-	-	-	-	✓	-
Checking key publications for terms and language used	-	-	✓	-	-	-	-	-	-	-
Internal validation in reference set independent from records used to derive filter terms										
	✓	X	N/A**	N/A**	X	X	X	✓	X	X
External validation in reference set independent from records used to derive filter terms and internal validation set										
	X	X	✓	✓	✓	X	X	✓	X	X

*Noel-Storr derived filter terms by running published search filters in MEDLINE combined with a subject search, locating 10 papers that all filters missed and choosing a term from the title/abstract or keywords of each.
** Only external validation was carried out (no internal validation) in real-world topics.
Abbreviations used: NR= not reported; N/A= not applicable

Table 2. Study characteristics and methods of MEDLINE development studies

Author (Year) Study ID	Identification of reference set	How was reference set used	How were search terms identified for filter	Ref set years	# gold standard records	# non-gold standard records	# journals ref set
Astin 2008	Hand search. Articles reporting on imaging as a diagnostic test in imaging journals. 6 high impact journals used to find studies for development set and 6 lower impact journals used to find studies for validation set. Journals indexed in MEDLINE and were also selected to cover general radiology, specific modalities and specific systems.	Two independent sets of records developed. Test set used to derive terms and test strategies. Validation set used to test external validity	Performed statistical analysis of terms in test set	development set 1985 Clin Radiol, 1988 Am J Neuroradiol; validation set 2000	333 in development set; 186 in validation set	2222 in development set; 1070 in validation set	12 (6 in development set; 6 in validation set)

Table 2. Study characteristics and methods of MEDLINE development studies *(Continued)*

Berg 2005	Manual review of a certain set of articles found using a search (via PubMed) combining sensitive terms for nursing literature plus cancer-related fatigue diagnosis terms. Manual review of these articles carried out to find diagnostic studies.	To derive terms and test strategies. Did not validate in a separate set of references	Existing PubMed Clinical Queries filter with extra terms from filters for CINAHL, medical publications, published recommendations & diagnosis definitions. Inductively collected terms derived from indexing of included citations: MeSH terms and frequently used text words in titles/abstracts.	NR	NR	238	NR
van der Weijden 1997	Personal literature database compiled over 10 years 'by every means of literature searching' of studies reporting on erythrocyte sedimentation rate as a diagnostic test.	To test strategies.	Checking key publications for definitions & terms used.	1985-1994	221	0	NR
Deville 2002	Studies included in two systematic reviews (relative recall).	To test strategies	Adapted three published search strategies	NR	NR	NR	NR
Deville 2000	Reference set of publications found through handsearch of 9 highest rank family medicine journals available on MEDLINE for years 1992-95. A 'control' set of publications for testing validity of strategies was found by adapting Haynes 1991 most sensitive and most specific searches by adding terms, then run in MEDLINE to retrieve all diagnostic primary studies, then limited to the 9 journals.	To derive terms from reference set; to test strategies in control set; to test external validity the best performing filters were compared against Haynes filters in a systematic review (SR) of meniscal lesions in the knee.	Performed statistical analysis of terms in reference set. Univariate analysis to calculate sensitivity, specificity & diagnostic odds ratio (DOR) of all relevant MeSH terms & text words. Models developed by forward stepwise logistic regression analysis.	1992-1995	75; 33 in meniscal lesions set	2392; NR meniscal lesions set	9
Haynes 2004	Manual review of 161 journals indexed on MEDLINE for year 2000. Journal titles regularly reviewed for appraisal for Evidence Based Med-	Test strategies and validate. The reference standard could not be	MeSH terms and text words listed using expert knowledge of the field.	2000	147	48881	161

Table 2. Study characteristics and methods of MEDLINE development studies *(Continued)*

	icine, Evidence Based Nursing, Evidence Based Mental Health and ACP Journal Club.	divided into a test set and validation set.					
Haynes 1994	Manual review of 10 high impact journals for the years 1986 and 1991. The 10 journals searched were American Journal of Medicine, Annals of Internal Medicine, Archives of Internal Medicine, BMJ, Circulation, Diabetes Care, Journal of Internal Medicine, JAMA, Lancet and NEJM	To test strategies and validate.	MeSH terms and text words listed using expert knowledge of the field.	1986 and 1991	92 in 1986 set; 111 in 1991 set	426 in 1986 set; 301 in 1991 set.	10
Bachmann 2002	Hand search European Journal of Paediatrics, Gastroenterology, American Journal of Obstetrics and Gynecology, and Thorax for years 1989 and 1994. Four different journals searched in 1999: NEJM, JAMA, BMJ and Lancet.	1989 set search used to derive terms and test strategies, 1994 and 1999 sets used to validate	Word frequency analysis on titles, abstracts and subject indexes of all references in 1989 set.	1989, 1994 and 1999	83 in 1989 test set; 53 in 1994 validation set; 61 in 1999 validation set.	1646 in 1989 test set; 1744 in 1994 validation set; 7875 in 1999 validation set	8
Vincent 2003	SRs retrieved from MEDLINE and EMBASE on OVID reporting on diagnostic tests for DVT. 16 SRs selected and all articles included that were indexed on MEDLINE became the reference set. Only English language articles included	To test strategies	Adapted from 5 published strategies: CASP, PubMed, Rochester, Deville, and North Thames	1969-2000	126	0	NR
Noel-Storr 2011	SR on the volume of evidence in biomarker studies in those with mild cognitive impairment, conducted by the authors.	To derive terms; to test strategies	Published search filters applied in MEDLINE combined with a subject search (Southampton A, Van der Weijden, and Southampton E), 10 papers were missed by all filters. One term from the title/abstract or keywords of each of 10 papers combined in the new filter.	2000-2011	128 in Sept 2010 set; additional 16 found in update search therefore 144 in August 2011	17266 in Sept 2010 set; additional 1654 found in update search therefore 18920 in August 2011	NR

Abbreviations used: NR=Not reported; ref set= reference set

Table 3. Performance of diagnostic filters from MEDLINE development studies

Author	Filter Description	Interface	Reference set	Sensitivity % (95% CI)	Specificity % (95% CI)	Accuracy (95% CI)	Precision% (95% CI)	NNR (95% CI)
Astin 2008	1. Exp "sensitivity and specificity"/ 2. False positive reactions/ 3. False negative reactions/ 4. du.fs 5. sensitivity.tw 6. (predictive adj4 value$).tw 7. distinguish$.tw 8. differentiat$.tw 9. enhancement.tw 10. identif$.tw 11. detect$.tw 12. diagnos$.tw 13. accura$.tw 14. comparison.tw 15. or/1-14	Ovid	Derivation set	95.8 (93.1, 97.5)	52.3 (50.2, 54.3)		23.1 (21.0-25.4)	0.04*
			Validation set	96.8 (93.1, 98.5)	43.9 (41.0, 46.9)		23.1 (20.3-26.2)	0.04*
Berg 2005	Some search terms were combined using "OR" thus increasing sensitivity and reducing specificity (e.g. nursing assessment [MeSH: noexp] AND questionnaire [Text Word]) Exemplary MeSH terms - Diagnosis, Differential; psychological tests; Likelihood functions; Area Under Curve; diagnostic tests; routine; diagnosis [MeSH subheading]; Diagnostic Techniques and Procedures; nursing assessment. Exemplary text words: sensitivity; specificity; predictive value; validity; reliability; likelihood ratio; questionnaire.	PubMed		87	73		Positive likelihood ratio (PLR)=3.2	2.3
	Some search terms were combined using "AND" thus increasing specificity and reducing sensitivity (e.g. nursing assessment [MeSH: noexp] AND questionnaire [Text Word]) Exemplary MeSH terms - Diagnosis, Differential; psychological tests; Likelihood func-	PubMed		76	83		PLR= 6.3	1.7

Table 3. Performance of diagnostic filters from MEDLINE development studies *(Continued)*

	tions; Area Under Curve; diagnostic tests; routine; diagnosis [MeSH subheading]; Diagnostic Techniques and Procedures; nursing assessment. Exemplary text words: sensitivity; specificity; predictive value; validity; reliability; likelihood ratio; questionnaire.						
Haynes 2004	sensitiv:.mp OR diagnos:.mp OR di.fs	Ovid	98.6 (96.8-100)	74.3 (73.9-74.7)	74.3 (74.0-74.7)	1.1 (1.0-1.3)	0.9*
	High specificity: specificity.tw	Ovid	64.6 (56.9-72.4)	98.4 (98.2-98.5)	98.3 (98.1-98.4)	10.6 (8.6-12.6)	0.09*
	High Sensitivity: di.xs.	Ovid	91.8 (87.4-96.3)	68.3 (67.9-68.7)	68.4 (68.0-68.8)	0.9 (0.7-1.0)	1.11*
	sensitiv:.mp OR predictive value:.mp OR accurac:.tw	Ovid	92.5 (88.3-96.8)	92.1 (91.8-92.3)	92.1 (91.8-92.3)	3.4 (2.8-3.9)	0.29*
	Optimising sensitivitiy and specificity: exp "diagnostic techniques and procedures"	Ovid	66.7 (59.1-74.3)	74.6 (74.2-75.0)	74.5 (74.2-74.9)	0.8 (0.6-0.9)	1.25*
	Sensitive:.mp. OR diagnos:.mp. OR accuracy.tw.	Ovid	98.0 (95.7-100.0)	82.7 (82.4-83.1)	82.8 (82.5-83.1)	1.7 (1.4-2.0)	0.59*
	Sensitive:.mp. OR diagnos:mp. OR test:.tw.	Ovid	98.0 (95.7-100.0)	75.1 (74.8-75.5)	75.2 (74.8-75.6)	1.2 (1.0-1.4)	0.83*
	Specificity.tw. OR predictive value:.tw.	Ovid	72.8 (65.6-80.0)	97.9 (97.8-98.1)	97.9 (97.7-98.0)	9.6 (7.9-11.3)	0.10*
	Accuracy:.tw. OR predictive value:tw.	Ovid	52.4 (44.3-60.5)	97.9 (97.8-98.1)	97.8 (97.7-97.9)	7.1 (5.6-8.6)	0.14*

Table 3. Performance of diagnostic filters from MEDLINE development studies *(Continued)*

	Sensitive:.mp. OR diagnostic.mp. OR predictive value:.tw.	Ovid		92.5 (88.3-96.8)	91.8 (91.6-92.1)	91.8 (91.6-92.1)	3.3 (2.8-3.8)	0.30*
	Exp sensitivity and specificity OR predicitive value:.tw.	Ovid		79.6	94.9	94.8	4.5	0.22*
Haynes 1994	*Best sensitivity:* diagnosis (subheading pre-explosion) OR specificity (tw)	NR		86	73	73	7	0.14*
	Best accuracy: Exp sensitivity and specificity OR diagnosis (subheading) OR diagnostic use (subheading) OR specificity (tw) OR predicitive (tw) AND value (tw))	NR		86	84	84	13	0.08*
	Best specificity: specificity (tw) OR (predictive (tw) AND value (tw)) OR (false (tw) and positive (tw))	NR		49	98		36	0.03*
	Best specificity: Exp sensitivity and specificity OR predictive (tw) AND value (tw)	NR		55	98		40	0.03*
	Diagnosis (subheading pre-explosion) OR Specificity (tw)	NR		86	73		7	0.14*
	Best sensitivity: Exp sensitivity and specificity OR diagnosis (subheading pre-explosion) OR diagnostic use (subheading) OR sensitivity (tw) OR specificity (tw)	NR		92	73		9	0.11*
			Haynes (2004) 2000 ref set	96.6	65	0.008	65.7	0.02*
	Diagnostic use (sh)	NR	1986 set	16	96		10	0.10*
			1991 set	26	96		18	0.06*
	Diagnosis (sh)	NR	1986 set	62	89		9	0.11*

Table 3. Performance of diagnostic filters from MEDLINE development studies *(Continued)*

			1991 set	59	88	13	0.08*
	Diagnosis& (px)	NR	1986 set	79	74	60	0.02*
			1991 set	80	77	90	0.01*
	Exp Sensitivity and Specificity	NR	1991 set	50	98	3	0.33*
	Specificity (tw)	NR	1991 set	54	96		
	Sensitivity (tw)	NR	1991 set	57	97		
			1986 set	43	98	3	0.33*
van der Weijden 1997	*MeSH short strategy (terms OR'd together)* explode DIAGNOSIS/diagnosis DIAGNOSIS-DIFFERENTIAL/all subheadings. explode SENSITIVITY-AND-SPECIFICITY REFERENCE-VALUES/all subheadings . FALSE-NEGATIVE-REACTIONS/ all subheadings . FALSE-POSITIVE-REACTIONS/ all subheadings . explode MASS-SCREENING/ all subheadings .	OVID		31		34	0.03*
	MeSH extended strategy (terms OR'd together) explode DIAGNOSIS/ all subheadings . explode SENSITIVITY-AND-SPECIFICITY REFERENCE-VALUES/all subheadings . FALSE-NEGATIVE-REACTIONS/ all subheadings . FALSE-POSITIVE-REACTIONS/ all subheadings . Explode MASS-SCREENING/ all subheadings .	OVID		69		11	0.09*
	MeSH extended and free text strategy explode DIAGNOSIS/ all subheadings . explode SENSITIVITY-AND-SPECIFICITY REFERENCE-VALUES/all subheadings . FALSE-NEGATIVE-REACTIONS/ all subheadings .	OVID		91		10	0.1*

Table 3. Performance of diagnostic filters from MEDLINE development studies *(Continued)*

	FALSE-POSITIVE-REACTIONS/ all subheadings . Explode MASS-SCREENING/ all subheadings . diagnos* OR sensitivity or specificity OR predictive value* OR reference value* OR ROC* OR likelihood ratio* OR monitoring					
Deville 2002	Sensitivity and specificity [Mesh; exploded] OR mass screening [Mesh; exploded] OR reference values [Mesh] OR false positive reactions [Mesh] OR false negative reactions [Mesh] OR specificit$.tw OR screening.tw OR false positive$.tw OR false negative$.tw	NR	Knee lesions SR	70		
			Urine dipstick SR	92		
Bachmann 2002	"SENSITIVITY AND SPECIFICITY"# OR predict* OR diagnos* OR sensitiv*	Datastar	1989 test set	92.8 (84.9-97.3)	15.6	6.4 (5.2-8.0)
			1994 validation set	98.1	10.9	9.2
			1999 validation set	91.8	4.7	21.3
	"SENSITIVITY AND SPECIFICITY"# OR predict* OR diagnos* OR accura*	Datastar	1989 test set	95.2 (88.1-98.7)	16.9	5.9 (4.8-7.3)
			1994 validation set	98.1 (89.9-99.9)	12 (9.1-1.4)	8.3 (6.7-11.3)
			1999 validation set	95.1	5	20.0
Vincent 2003	*Strategy A* 1. exp 'sensitivity and specificity'/; 2. (sensitivity or specificity or accuracy).tw.; 3. ((predictive adj3 value$) or (roc adj curve $)).tw.; 4. ((false adj positiv$) or (false negativ$)).tw.;	Ovid		100	3*	0.33*

Table 3. Performance of diagnostic filters from MEDLINE development studies *(Continued)*

5. (observer adj variation$) or (likelihood adj3 ratio$)).tw.; 6. likelihood function/; 7. exp mass screening/; 8. diagnosis, differential/ or exp Diagnostic errors/; 9. di.xs or du.fs; 10. or/1-9				
Strategy B 1. exp 'sensitivity and specificity'/; 2. (sensitivity or specificity or accuracy).tw.; 3. (predictive adj3 value$); 4. exp Diagnostic errors/; 5. ((false adj positiv$) or (false adj negativ$)).tw; 6. (observer adj variation$).tw; 7. (roc adj curve$).tw; 8. (likelihood adj3 ratio$).tw.; 9. likelihood function/; 10. exp *venous thrombosis/di, ra, ri, us; 11. exp *thrombophlebitis/di, ra, ri, us; 12. or/1-11	Ovid	98.4	5*	0.2*
Strategy C 1. exp 'sensitivity and specificity'/; 2. (sensitivity or specificity or accuracy).tw.; 3. ((predictive adj3 value$) or (roc adj curve $)).tw.;	Ovid	79.4	10*	0.1*

Table 3. Performance of diagnostic filters from MEDLINE development studies *(Continued)*

	4. ((false adj positiv$) or (false negativ$)).tw.; 5. (observer adj variation$); 6. likelihood function/ or; 7. exp Diagnostic errors/; 8. (likelihood adj3 ratio$).tw.; 9. or /1-8						
Deville 2000	*Strategy 4* SENSITIVITY AND SPECIFICITY (exp) OR specificity (tw) OR false negative (tw) OR accuracy (tw) OR screening (tw)	NR		89.3 (82.3-96.3)	91.9 (90.8-93)		DOR 95
			Meniscal lesion	61 (42.1-77.1)		4.7	0.22*
	Strategy 3 SENSITIVITY AND SPECIFICITY (exp) OR specificity (tw) OR false negative (tw) OR accuracy (tw)	NR		80.0 (71.0-89.1)	97.3 (96.6-97.9)	48 (40-56)	DOR 149
	Strategy 2 SENSITIVITY AND SPECIFICITY (exp) OR specificity (tw) OR false negative (tw)	NR		73.3 (63.3-83.3)	98.4 (97.9-98.9)		DOR 170
	Strategy 1 SENSITIVITY AND SPECIFICITY (exp) OR specificity (tw)	NR		70.7 (60.4-81.0)	98.5 (98.0-98.9)		DOR 158
Noel-Storr 2011	1. Disease progression/ 2. di.fs. 3. logitudinal*.ab. 4. Follow-up studies/ 5. conversion.ab. 6. transition.ab.	Ovid	2000-Sept 2010	97 (92-99)	38 (37-39)	1.1 (0.95-1.4)	
			2000-Aug 2011	98 (94-100)	38 (37-39)	1.2 (1.0-1.4)	

Table 3. Performance of diagnostic filters from MEDLINE development studies *(Continued)*

7. converters.ab.

8. progressive.ab.

9. "increased risk".ab.

10. "follow-up".ab.

NR=Not reported

Table 4. Summary of study design characteristics of MEDLINE filter evaluation studies

	Evaluation study: Author (year)									
	Kastner 2009	Ritchie 2007	Leeflang 2006	Kassai 2006	Doust 2005	Whiting 2010	Vincent 2003	Deville 2000	Mitchell 2005	Noel-Storr 2011
Method of identification of reference set records(one from list below selected for each study)										
• Handsearching for primary studies	-	-	-	-	-	-	-	✓	✓	-
• Internet search for DTA systematic reviews	✓	-	✓	-	-	-	✓	-	-	-
• Systematic reviews conducted by authors	-	✓	-	-	✓	✓	-	-	-	✓
• Primary studies identified through Internet search				✓	-	-	-	-	-	-
If systematic reviews used in reference set development, did they include DTA search terms in search strategy?	✓	X	Unclear	-	✓	X	✓	-	-	X
Reference set also contained non-gold standard records	X	✓	X	✓	X	✓	X	✓	✓	✓
Description of non-gold standard records if contained in reference set	-	-	-	-	-	NR	-	-	NR	-

Table 4. Summary of study design characteristics of MEDLINE filter evaluation studies *(Continued)*

• All studies retrieved by search not classified as gold-standard records	-	✓	-	✓	-	-	-	-	-	✓
• False positive papers selected by a previously published search strategy, exclusion of some publication types e.g. reviews and meta-analyses.	-	-	-	-	-	-	-	✓	-	-
Generic gold standard records I.e. not topic specific	✓	X	✓	X	X	✓	✓	X	X	X

Table 5. Study characteristics and methods of included MEDLINE filter evaluation studies

Study	Identification of reference set	Reference set selection criteria	Ref set years	# gold standard records	# non-gold standard records	# journals ref set if hand-searched	Definition of DTA study if hand-searched gold standard identified	Description of filter allows reproducibility	Definitions of Se & Sp	Number of filters evaluated
Kastner 2009	Included studies from 12 published SRs on the ACP Journal Club website and indexed on MEDLINE or EMBASE.	Eligibility criteria for including SR were: published in 2006; incorporated a MEDLINE and EMBASE search as a data source; and available and downloadable in electronic format. In addition the review cannot have used the Clinical Queries filter, but other search filters were permissible.	2006 (date publication SRs	441	0	Not given. The 12 SRs were from 9 journals.	The study compared at least two diagnostic test procedures with one another.	yes	no	1

Table 5. Study characteristics and methods of included MEDLINE filter evaluation studies *(Continued)*

Ritchie 2007	SR of DTA studies for UTI in young children carried out by the authors	Included studies that could be identified in Ovid MEDLINE	1966-2003	160	27804	NA	NA	no	no	22
Leeflang 2006	Included studies from 27 published SR. Reviews selected after an electronic search for SRs of DTA studies published between January 1999 and April 2002 in MEDLINE, EMBASE, DARE and Medion	Criteria for inclusion SRs: assessment of DTA, the inclusion of >10 original studies with inclusion not based on design characteristics, and sufficient data to reproduce the contingency table. Exclusion of reviews that reported the application of a diagnostic search filter.	1999-2002	820	0	NA	NA	yes	no	12
Kassai 2006	Used PubMed interface to search MEDLINE, Science Citation Index, EMBASE and Pascal Biomed for relevant articles using search strategies with terms (MeSH and free text for MEDLINE) related to venous thrombosis, venography and ultrasonography in all databases.	Any relevant article retrieved through topic search on MEDLINE, Science Citation Index, EMBASE and Pascal Biomed	1966-2002	237	1236	NR	NR		yes	3
Doust 2005	Included studies from two SRs: tympanometry (TR) for the diagnosis of otitis media with effusion in children, and natriuretic peptides (NPR). Initial list of citations was generated from MEDLINE us-	Included in two SRs conducted by the authors	TR 1966-2001; NPR 1994-2002	TR n=33; NPR n=20	TR n=0; NPR n=0	TR n=22; NPR n=16	NR	yes	yes	5

Table 5. Study characteristics and methods of included MEDLINE filter evaluation studies *(Continued)*

	ing the search strategy used by the sensitivity option of the Clinical Queries filter for DTA in PubMed. Reference lists of potentially relevant papers and review articles were checked for further possible papers.									
Whiting 2010	Test accuracy studies indexed on MEDLINE from 7 SRs carried out by authors. Relative recall reference set.	All included studies indexed on MEDLINE from 7 SRs of DTA. SRs that conducted extensive searches that were not limited using methodological filters or search terms relating to measures of test accuracy	NR	506	25880**	NR	Studies in which cross-tabulation data comparing the results of the index test with the reference standard were available.	yes	yes	22
Vincent 2003	SRs retrieved from MEDLINE and EMBASE on OVID using validated SR filter on diagnostic tests for DVT. 16 SR selected and all articles included that were indexed on MEDLINE became the reference set. Only English language articles included	Studies included in 16 SRs that compared one of the specified diagnostic tests for DVT against a venogram.	1969-2000	126	0	NR	Compared specified diagnostic test to reference standard	yes	yes	5
Deville 2000	Adapted Haynes 1991 most sensitive and specific filter by adding terms. Ran search in MEDLINE to retrieve all primary DTA studies. Second	Primary DTA studies indexed on MEDLINE; studies included on physical tests for the diagnosis of meniscal lesions of the knee.	1992-1995	75; 33 in meniscal lesions set	2392; NR in meniscal lesions set		Diagnostic test was compared with a reference standard	yes	yes	1

Table 5. Study characteristics and methods of included MEDLINE filter evaluation studies *(Continued)*

	set of references selected on diagnosis of meniscal lesions of the knee for external validity testing. No further details on how this set was selected are provided.									
Mitchell 2005	Handsearch of the 3 top ranking renal journals for the years 1990-1991 and 2002-2003.	Primary DTA studies that could be identified in MEDLINE on the diagnosis of kidney disease	1991-1992 2002-2003	99	4409	3	A test or tests being compared to a reference standard in a human population	yes*	NR	6
Noel-Storr 2011	SR on the volume of evidence in biomarker studies in those with mild cognitive impairment, conducted by the authors.	Primary DTA longitudinal studies indexed on MEDLINE with at least one follow-up period; at least one of biomarkers of interest used as test of interest; included subjects with objective cognitive impairment at baseline, no dementia.	2000-Sept 2010; 2000-Aug 2011	128 Sept 2010; 144 Aug 2011	17266 Sept 2010; 18920 Aug 2011	NR	NA	yes*	no	22

Abbreviations used: TR= Tympanometry review; NPR= Natriuretic peptides review; SR= systematic review; NA= not applicable; NR= not reported; ref set= reference set; Se= sensitivity; Sp= specificity.
* Full strategy obtained from authors
** Number of gold-standard records obtained from authors

Table 6. Summary of study design characteristics of EMBASE filter development studies

	Author	
	Bachmann 2003	Wilczynski 2005
Method of identification of reference set records (one from list below selected for each study)		
• Hand-searching for primary studies	✓	✓
• DTA systematic reviews	-	-
• Personal literature database	-	-
Reference set also contained non-gold standard records	✓	✓
Description of non-gold standard records if contained in reference set	-	NR
• All records retrieved by search that were not classified as gold standard studies	✓	
Generic gold standard records i.e. not topic specific	✓	✓
Method of deriving filter terms (a combination of methods could be used)		
• Analysis of reference set	✓	✓
• Expert knowledge	-	✓
• Adaption of existing filter	-	-
• Checking key publications for terms and language used	-	-
Internal validation in reference set independent from records used to derive filter terms	x	X
External validation in reference set independent from records used to derive filter terms and internal validation set	x	x

Abbreviations used: NR= not reported

Table 7. Study characteristics and methods of EMBASE filter development studies

Study	Identification of reference set	How was reference set used	How were search terms identified for filter	Ref set years	# gold standard records	# non-gold standard records	# journals ref set
Bachmann 2003	Handsearching of all issues of NEJM, Lancet, JAMA and BMJ published in 1999.	To derive terms; to test strategies	Word frequency analysis on title, abstract and subject indexing of handsearched records	1999	61	6082	4
Wilczynski 2005	Handsearching each issue of 55 journals in 2000.	To test strategies	Initial list of MeSH terms and text words compiled using knowledge of the field and input from librarians and clinicians. Stepwise logistic regression used to improve performance of filters.	2000	97	27,672	55

Abbreviations used: ref set= reference set

Table 8. Summary of study design characteristics of EMBASE filter evaluation studies

	Author		
	Kastner 2009	Wilczynski 2005	Mitchell 2005
Method of identification of reference set records(one from list below selected for each study)			
• Handsearching for primary studies	-	✓	✓
• Internet search for DTA systematic reviews	✓	-	-
• Systematic reviews conducted by authors	-	-	-
• Primary studies identified through Internet search		-	-
If systematic reviews used in reference set development, did they include DTA search terms in search strategy?	✓	-	-
Reference set also contained non-gold standard records	x	✓	✓
Description of non-gold standard records if contained in reference set	NR	NR	NR
Generic gold standard records i.e. not topic specific	✓	✓	x

Abbreviations used: NR= not reported

Table 9. Study characteristics and methods of studies evaluating EMBASE filters

Study	Identification of gold standard	Reference set selection criteria	Ref set years	# gold standard studies ref set	# non-gold standard studies in ref set	# journals ref set for hand-searched gold standard	Definition of DTA study	Description of filter allows reproducibility	Definitions of Se & Sp	Number of filters evaluated
Kastner 2009	Included studies from 12 published SRs on the ACP Journal Club website and indexed in MEDLINE or EMBASE.	Eligibility criteria for including SR were: published in 2006; incorporated a MEDLINE and EMBASE search as a data source; and available and downloadable in electronic format. In addition the review cannot have used the Clinical Queries filter.	2006 (date SRs published)	441	441	NA	The study compared at least two diagnostic test procedures with one another.	yes	no	1
Wilczynski 2005	Handsearch of each issue of 55 journals in 2000.	Studies indexed in EMBASE found through handsearching which met the methodological criteria for a diagnostic study:	2000	97	27575	55	Inclusion of spectrum of participants; reference standard; participants received both the new test of reference standard; interpretation of index test without knowledge of reference standard and vice versa; analysis consistent with study design.	yes	yes	2
Mitchell 2005	Handsearch of the 3 top ranking renal journals for the years 1990-1991	Primary DTA studies that could be identified in EMBASE reporting on the accuracy of tests for kidney disease diagnosis	1991-1992 2002-2003	96	3984	3	A test or tests being compared to a reference standard in a human population	yes*	no	4

Table 9. Study characteristics and methods of studies evaluating EMBASE filters *(Continued)*

and
2002-2003

Abbreviations used: ref set= reference set; Se= sensitivity; Sp= specificity

Table 10. MEDLINE filters evaluated by two or more studies (values given in percentages)

SENSITIVITY											SPECIFICITY				PRECI
ORIGINAL DEVELOPMENT STUDY	WHITING	LEEFLANG	KASTNER	*DOUSTTR	*DOUSTNPR	VINCENT	DEVILLE	DEVILLE ML	KASSAIMITCHELL	NOEL-STORR	**ORIGINAL DEVELOPMENT STUDY**	WHITING	MITCHELL	NOEL-STORR	**ORIGINAL DEVELOPMENT STUDY**
Original development study did report performance data															
95 Wilczynski 2004 Sensitive	87	88		70	90				84	84	**NR**	37	80	36	**5.0**
99 Haynes 2004 Sensitive	80	87	88	70	100				67	69	**74**	41	85	45	**1.1**
65 Haynes 2004 Specific	43	28								14	**98**	94		95	**10.6**
ML=89 Deville 2000 Strategy	68	46		58	100	75			49	55	**92**	81	95	82	**NR**
ML=61															**4.7**

Table 10. MEDLINE filters evaluated by two or more studies (values given in percentages) *(Continued)*

gy 4																	
55 Haynes 1994 Specific	55	29									51	**98**	90			88	**40.0**
72 Haynes 1994 Sensitive**	85	81			96	73	45	95	80		91	**73**	23		80	32	**9.0**
77 Vincent 2003 Strategy C	67	44									54	**NR**	85			83	**10.0**
91 van der Weijden 1997 Sensitive	87	92	73	100					96		93	**NR**	15		96	30	**NR**
USR=70 Ville 2002 Accurate **USR=92**		51															
86 Haynes 1994 Accurate		81										**84**					**13.0**

Table 10. MEDLINE filters evaluated by two or more studies (values given in percentages) *(Continued)*

cu-rate																		
80 Nev-ille 2000 Strat-e-gy 3	41													**97**				**48.0**
71 Nev-ille 2000 Strat-e-gy 1										76				**99**				
100 cent 2003 Strat-e-gy A											81			**NR**			81	**2.5**
Original development study did NOT report performance data																		
NR ck-Yt-ter 2004**	85													71		**NR**	39	51	**NR**
NR 2002 $	83					100				95				67		**NR**	53	49	**NR**
NR Nev-ille 2002a Ex-	71			58		100								60		**NR**	78	78	**NR**

Table 10. MEDLINE filters evaluated by two or more studies (values given in percentages) *(Continued)*

tend- ed							
NR	NR	86	87	NR	39	33	NR
erdeen In- ter- TASC 2011 $							
NR	NR	86	93	NR	13	29	NR
ton A In- ter- TASC 2011 $							
NR	NR	69	55	NR	80	81	NR
ton B In- ter- TASC 2011 $							
NR	NR	56	51	NR	90	88	NR
ton C In- ter- TASC 2011							
NR	NR	84	89	NR	21	42	NR
ton D							

Table 10. MEDLINE filters evaluated by two or more studies (values given in percentages) *(Continued)*

Inter-TASC 2011							
NR Southampton E Inter-TASC 2011 $	87	92	NR	14	31	NR	
NR SIGN A Inter-TASC 2011	73	70	NR	62	58	NR	
NR SIGN B Inter-TASC 2011	64	67	NR	81	71	NR	
NR SIGN C Inter-TASC 2011	85	90	NR	24	43	NR	
NR NHS-BS Inter-	69	56	NR	83	80	NR	

Table 10. MEDLINE filters evaluated by two or more studies (values given in percentages) *(Continued)*

TASC 2011								
NR Shipley Miner 2002	72			63	NR	73	73	NR
NR Deville 2002a Accurate	88				NR			NR
NR University of Rochester 2002 $		79			NR			NR
NR North Thames 2002 $		53			NR			NR

* Doust combines each methodological filter with a content filter for a Tympanometry systematic review (TR) and a Natriuretic peptides systematic review (NPR), this is the reason for two results being reported for each filter.
Similarly, Deville (2000) uses an independent set of references to externally validate their own filter and the Haynes 1994 sensitive filter; ALL= all references in main reference set; ML= references on the diagnosis of meniscal lesions of the knee.
** Falck-Ytter filter is an adaption of the Hanyes 1994 sensitive filter for OVID into a PubMed format (alternative to the PubMed Clinical Queries adaption of the same filter).
Abbreviations used: KSR= Knee lesion systematic review; USR= Urine dipstick systematic review.
$ Filter no longer available from source cited by evaluation studies.

Table 11. Performance of EMBASE filters from development studies

Author (year) ID	Filter Description	Filter interface	Sensitivity % (95% CI)	Specificity % (95% CI)	Precision % (95% CI)	NNR
Bachmann 2003	sensitiv* OR detect* (specific filter)	Datastar, Ovid and Silverplatter	73.7 (60.9-84.2)		17.6	5.7 (4.4-7.6)
	sensitiv* OR detect* OR accura* OR specific* OR reliab* OR positive OR negative OR diagnos*	Datastar, Ovid and Silverplatter	100 (94.1-100)		3.7	27.0 (21.0-34.8)
	sensitiv* OR detect* OR accura*	Datastar, Ovid and Silverplatter	85.2		14.2	7.0
	sensitiv* OR detect* OR accura* OR specific*	Datastar, Ovid and Silverplatter	86.9		10.4	9.6
	sensitiv* OR detect* OR accura* OR specific* OR reliab*	Datastar, Ovid and Silverplatter	90.2		10.4	9.6
	sensitiv* OR detect* OR accura* OR specific* OR reliab* OR positive	Datastar, Ovid and Silverplatter	91.8		9.2	10.9
	sensitiv* OR detect* OR accura* OR specific* OR reliab* OR positive OR negative	Datastar, Ovid and Silverplatter	91.8		8.5	11.8
	sensitiv*	Datastar, Ovid and Silverplatter	45.9		27.7	3.6
Wilczynski 2005	Best sensitivity: di.fs OR predict:.tw OR specificity.tw	Ovid	100 (100-100)	70.4 (69.8-70.9)	1.2 (0.9-1.4)	
	Small drop in sensitivity with substantive gain in specificity: diagnos:.mp OR predict:.tw OR specificity.tw	Ovid	96.9 (93.5-100)	78.2 (77.7-78.7)	1.5 (1.2-1.8)	
	Small drop in specificity with a substantive gain in sensitivity: specificity.tw OR accurac:.tw	Ovid	73.2 (64.4-82.0)	97.4 (97.2-97.5)	8.8 (6.9-10.8)	
	Best optimal strategy: sensitiv:.tw OR diagnostic accuracy.sh OR diagnostic.tw	Ovid	89.7 (83.6-95.7)	91.6 (91.3-91.9)	3.3 (2.9-4.4)	

Table 12. Performance of evaluated EMBASE filters

Filter (original reference)	Author (year) of evaluation study	Description of filter from evaluation paper	Interface filter developed for	Sensitivity %	Precision %	Comments and other measures

Table 12. Performance of evaluated EMBASE filters *(Continued)*

PubMed Clinical Queries Ovid 2010	**ORIGINAL**	sensitiv:.mp. OR diagnos:.mp. OR di.fs.	Ovid	NR	NR	
	Kastner 2009	sensitiv:.mp. OR diagnos:.mp. OR di.fs.	Ovid	88		
Bachmann 2003 Sensitive	**ORIGINAL**	sensitiv* OR detect* OR accura* OR specific* OR reliab* OR positive OR negative OR diagnos*	Datastar, Ovid and Silverplatter	100	3.7	
	Wilczynski 2005	sensitiv:.tw. OR detect:.tw. OR accura:.tw. OR specific:.tw. OR reliab:.tw. OR positive:.tw. OR negative:.tw. OR diagnos:.tw.	Ovid	97	1.2	Specificity=72.%; Accuracy=72.%
	Mitchell 2005	sensitive* OR detect* OR accura* OR specific* OR reliab* OR positive OR negative OR diagnos*	Ovid	86	4.4	Specificity=60%
Bachmann 2003 Specific	**ORIGINAL**	sensitiv* OR detect*	Datastar, Ovid and Silverplatter	74	17.6	NNR=5.7
	Mitchell 2005	sensitiv* .tw. OR detect* .tw.	Ovid	79	3.0	Specificity=91%; Accuracy=91%
Wilczynski 2005 Sensitive	**ORIGINAL**	di.fs OR predict:.tw OR specificity.tw	Ovid	100	1.2	Specificity=70%; Accuracy=71%
	Mitchell 2005	di.fs OR predict* .tw. OR specificity.tw.	Ovid	72	9	Specificity=83%

Abbreviations used: NR= not reported

APPENDICES

Appendix 1. MEDLINE search strategy

MEDLINE ® OvidSP 1950 to week 1 November 2012

1 “Information Storage and Retrieval”/

2 ((information or literature) adj5 retriev$).tw.

3 Databases, Bibliographic/

4 ((bibliographic adj1 database$) or (electronic adj1 database$) or (online adj1 database$)).tw.

5 Medline/

6 PubMed/

7 Medlars/

8 Grateful Med/

9 (medline or pubmed or medlars or grateful-med or gratefulmed or embase$ or excerpta medica).tw.

10 or/1-9

11 (search$ adj5 (strateg$ or filter$ or hedge$ or technique$ or term$1)).tw.

12 (retriev$ adj5 (strateg$ or filter$ or hedge$ or technique$)).tw.

13 ((methodology or methodologic$) adj5 (strateg$ or filter$ or hedge$ or search$ or term$1)).tw.

14 (search$ adj5 (precision or recall or accura$ or sensitiv*)).tw.

15 (retriev$ adj5 (precision or recall or accura$ or sensitiv$)).tw.

16 or/11-15

17 (diagnos$ adj5 (strateg$ or filter$ or hedge$ or search$ or term$1)).tw.

18 exp Diagnosis/

19 diagnos$.tw.

20 "Sensitivity and Specificity"/

21 (sensitiv$ and specific$).tw.

22 or/18-21

23 10 and 16 and 22

24 10 and 17

25 23 or 24

26 "cochrane database of systematic reviews".so.

27 25 not 26

Appendix 2. EMBASE search strategy

EMBASE OvidSP 1980 to 2012 Week 48

1 Information Retrieval/

2 ((information or literature) adj5 retriev$).tw.

3 Bibliographic Database/

4 ((bibliographic adj1 database$) or (electronic adj1 database$) or (online adj1 database$)).tw.

5 Medline/ or Embase/

6 (medline or pubmed or medlars or grateful-med or gratefulmed or embase$ or excerpta medica).tw.

7 or/1-6

8 (search$ adj5 (strateg$ or filter$ or hedge$ or technique$ or term$1)).tw.

9 (retriev$ adj5 (strateg$ or filter$ or hedge$ or technique$)).tw.

10 (search$ adj5 (precision or recall or accura$ or sensitiv*)).tw.

11 (retriev$ adj5 (precision or recall or accura$ or sensitiv$)).tw.

12 ((methodology or methodologic$) adj5 (strateg$ or filter$ or hedge$ or search$ or term$1)).tw.

13 or/8-12

14 (diagnos$ adj5 (strateg$ or filter$ or hedge$ or search$ or term$1)).tw.

15 exp "Diagnosis, Measurement and Analysis"/

16 diagnos$.tw.

17 "Sensitivity and Specificity"/

18 (sensitiv$ and specific$).tw.

19 or/15-18

20 7 and 13 and 19

21 7 and 14

22 20 or 21

23 "cochrane database of systematic reviews".so.

24 “cochrane database of systematic reviews (online)”.so.

25 23 or 24

26 22 not 25

Appendix 3. ISI Web of Science search strategy

ISI Web of Science searched 11 January 2013

ISI Web of Science Databases=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, CCR-EXPANDED, IC Timespan=All Years

6 #3 AND #4 AND #5

5 #1 OR #2

4 TS=diagnos*

3 TS=(information retriev* OR literature retriev* OR bibliographic database OR medline OR pubmed OR medlars OR grateful med OR gratefulmed OR embase* OR psycinfo)

2 TS=(retriev* same (strateg* OR filter* OR hedge* OR technique*))

1 TS=(search* same (strateg* OR filter* OR hedge* OR technique* OR term*))

Appendix 4. PsycINFO search strategy

PsycINFO (OvidSP) searched 13 March 2013

1. exp Automated Information Retrieval/

2. Databases/

3. Information Seeking/

4. Computer Searching/

5. ((information or literature) adj2 retriev$).tw.

6. ((bibliographic adj1 database?) or (electronic adj1 database?)).tw.

7. (medline or pubmed or medlars or grateful med or gratefulmed or embase$ or excerpta medica).tw.

8. psycinfo.ti.

9. psycinfo.ab. /freq=2

10. or/1-9

11. (search$ adj2 (strateg$ or filter$ or hedge? or technique? or term$1)).tw.

12. (retriev$ adj2 (strateg$ or filter$ or hedge? or technique?)).tw.

13. (sensitiv$ or specific$ or recall or precision or precise or number needed to read or NNR).tw.

14. or/11-13

15. Diagnosis/

16. diagnos$.tw.

17. or/15-16

18. and/10,14,17

Appendix 5. Library, Information Science and Technology Abstracts (LISTA) search strategy

Library, Information Science and Technology Abstracts (LISTA) strategy searched 13 March 2013

S41 S37 or S40

S40 S16 and S29 and S39

S39 S28 or S38

S38 S30 or S31 or S32 or S33 or S34 or S35

S37 S16 and S28 and S36

S36 S29 or S30 or S31 or S32 or S33 or S34 or S35

S35 NNR

S34 "number needed to read"

S33 precision

S32 recall

S31 specificity

S30 sensitivity

S29 diagnos*

S28 (S17 or S18 or S19 or S20 or S21 or S22 or S23 or S24 or S25 or S26 or S27)

S27 retriev* N2 techniqu*

S26 retriev* N2 hedge*

S25 retriev* N2 filter*

S24 retriev* N2 strateg*

S23 search* N2 terms

S22 search* N2 term

S21 search* N2 techniqu*

S20 search* N2 hedge*

S19 search* N2 filter*

S18 search* N2 strateg*

S17 DE Search Algorithms

S16 (S1 or S2 or S3 or S4 or S5 or S6 or S7 or S8 or S9 or S10 or S11 or S12 or S13 or S14 or S15)

S15 medline OR pubmed or medlars or "grateful med" or gratefulmed or embase* or "excerpta medica"

S14 DE Electronic Information Resources

S13 DE Bibliographic Databases

S12 DE Databases

S11 DE PubMed

S10 DE EMBASE

S9 DE MEDLINE

S8 DE "Information Storage & Retrieval Systems"

S7 information N2 search*

S6 literature N2 search*

S5 literature N2 retriev*

S4 information N2 retriev*

S3 DE "electronic information resource searching"

S2 DE "database searching"

S1 DE "information retrieval"

Appendix 6. Cochrane Methodology Register search strategy

Cochrane Methodology Register 2012, Issue 3 in *The Cochrane Library* (Wiley InterScience Online)

#1 ("diagnostic test accuracy" NEXT "search strategies"):kw in Methods Studies

#2 ("study identification" next general) or ("study identification" next "prospective registration") or ("study identification" next "internet") or ("information retrieval" next general) or ("information retrieval" next "retrieval techniques") or ("information retrieval" next "comparisons of methods") or ("information retrieval" next indexing):kw in Methods Studies

#3 search*:ti NEAR/5 (strateg* or filter* or hedge* or technique* or term or terms or precision or recall or accura*):ti in Methods Studies

#4 retriev*:ti NEAR/5 (strateg* or filter* or hedge* or technique* or term or terms or precision or recall or accura*):ti in Methods Studies

#5 search*:ab NEAR/5 (strateg* or filter* or hedge* or technique* or term or terms or precision or recall or accura*):ab in Methods Studies

#6 retriev*:ab NEAR/5 (strateg* or filter* or hedge* or technique* or term or terms or precision or recall or accura*):ab in Methods Studies

#7 methodology:ti NEAR/5 (strateg* or filter* or hedge* or term or terms):ti in Methods Studies

#8 methodologic*:ti NEAR/5 (strateg* or filter* or hedge* or term or terms):ti in Methods Studies

#9 methodology:ab NEAR/5 (strateg* or filter* or hedge* or term or terms):ab in Methods Studies

#10 methodologic*:ab NEAR/5 (strateg* or filter* or hedge* or term or terms):ab in Methods Studies

#11 (medline or pubmed or medlars or "grateful med" or gratefulmed or embase* or excerpta medica):ti in Methods Studies

#12 (medline or pubmed or medlars or "grateful med" or gratefulmed or embase* or excerpta medica):ab in Methods Studies

#13 (diagnos* or sensitiv* or specific*):ti in Methods Studies

#14 (diagnos* or sensitiv* or specific*):ab in Methods Studies

#15 (#2 OR #3 OR #4 OR #5 OR #6 OR #7 OR #8 OR #9 OR #10) AND (#11 OR #12)

#16 (#2 OR #3 OR #4 OR #5 OR #6 OR #7 OR #8 OR #9 OR #10) AND (#13 OR #14)

#17 diagnos*:ti NEAR/5 (strateg* or filter* or hedge* or search* or term or terms):ti in Methods Studies

#18 diagnos*:ab NEAR/5 (strateg* or filter* or hedge* or search* or term or terms):ab in Methods Studies

#19 (#17 OR #18)

#20 (#1 OR #15 OR #16 OR #19)

Appendix 7. Library and Information Science Abstracts (LISA) search strategy

LISA: Library and Information Science Abstracts (Cambridge Scientific Abstracts) - searched 31 May 2010

(((DE=("databases" or "bibliographic databases" or "cd rom
databases" or "database producers" or "online databases" or "computerized
information retrieval" or "multiple database searches" or "online
information retrieval")) or (TI=((literature or information) within 2
retriev*)) or (AB=((literature or information) within 2 retriev*))
or (TI=((bibliographic or electronic) within 2 database*))
or (AB=((bibliographic or electronic) within 2 database*)) or (TI=(medline
or medlars or pubmed or grateful med or gratefulmed or embase* or
excerpta medica)) or (AB=(medline or medlars or pubmed or grateful med or
gratefulmed or embase* or excerpta medica))) and ((DE=("search strategies"
or "searching" or "boolean strategies" or "non boolean strategies" or
"term selection" or "free text searching" or "full text searching" or
"ranking")) or (DE=("boolean strategies" or "non boolean strategies"))
or (TI=(search* within 2 (strateg* or filter? or hedge? or technique? or
term?))) or (AB=(search* within 2 (strateg* or filter? or hedge? or
technique? or term?))) or (TI=(retriev* within 2 (strateg* or filter? or
hedge? or technique?))) or (AB=(retriev* within 2 (strateg* or filter? or
hedge? or technique?)))))) and ((TI=diagnos* or AB=diagnos*)
or (DE=("recall" or "retrieval performance measures" or "exhaustivity" or
"pertinence" or "relevance")) or (DE="retrieval performance measures")
or (TI=(sensitivity or specificity or recall or precision or accuracy or
(number within3 read))) or (AB=(sensitivity or specificity or recall or
precision or accuracy or (number within3 read))))

Appendix 8. Definition of terms used in this review

Accuracy – proportion of all articles correctly categorised

Development study – a study which aims to develop and test a search strategy for locating diagnostic test accuracy studies

Diagnostic odds ratio – positive likelihood ratio/negative likelihood ratio

Diagnostic test accuracy study – a study which compares the results of the test of interest, the index test, to those of a reference standard, which should be the best available method of determining disease status

Evaluation study – a study which quantitatively evaluates existing search strategies for locating diagnostic test accuracy studies

Gold standard record – a record included in the reference set that meets the criteria for a diagnostic test accuracy study

Non-gold standard record – a record included in the reference set that does not meet the criteria for a diagnostic test accuracy study

Number Needed to Read – the number of articles needed to read to identify one relevant article, calculated as 1 divided by precision

Positive likelihood ratio – the proportion of the probability of a true positive record to the false positive records

Precision/positive predictive value – proportion of retrieved records meeting diagnostic test criteria – proportion of gold standard records in the result set

Reference set – compilation of records which can be used to derive terms for search filter development and test the performance of search filters. The reference set can be composed of gold standard and non-gold standard records, or gold standard records alone

Sensitivity – percentage of correctly identified gold standard studies

Specificity – percentage of correctly non-identified studies.

	Reference set	
Search terms	**Gold standard records**	**Non-gold standard**
Detected	a	b
Not detected	c	d

Sensitivity = a/(a + c); precision = a/(a + b); specificity = d/(b + d); accuracy = (a + d)/(a + b + c + d). All included and excluded references in gold standard = (a + b + c + d)

Appendix 9. MEDLINE filters with full strategies as used by evaluation studies

Filter (original reference)	**Author (year) of evaluation study**	**Description of filter as appears in evaluation study**	**Interface used by evaluation study**	**Sensitivity (95% CI)**	**Precision (95% CI)**	**Comments**
Bachmann 2002	**ORIGINAL**	exp sensitivity and specificity or predict$ or diagnos$ or di.fs. or du.fs. or accura$				
Sensitive	Ritchie 2007	NR	Ovid	74	1.36	

(Continued)

	Leeflang 2006	"sensitivity and specificity"[MeSH] OR predict* OR diagnose* OR diagnosi* OR diagnost* OR accura*	PubMed	88		
	Doust 2005	Sensitivity and specificity [MeSH] predict* [tw] diagnos* [tw] accura* [tw]	Datastar, Ovid, PubMed, Silverplatter	70	5	Methodological & content filter for TSR
				90	4	Methodological & content filter for NPSR
				88		Methodological filter for TSR
				90		Methodological filter for NPSR
	Whiting 2010	Exp "sensitivity and specificity"/ Diagnos$ OR di.fs. or du.fs. Predict$ Accura$	Ovid	87 (81-98)	3 (1-22)	NNR 36 (4-98)
	Noel-Storr 2011	NR	Ovid	84 (77-90)	0.17 (0.14-0.20)	
	Mitchell 2005	1. exp "Sensitivity and Specificity"/ 2. predict$.tw. 3. diagnos$.tw. 4. di.fs. 5. du.fs. 6. accura$.tw. 7. or/1-6	Ovid	84	8.8	Strategy from Table 3
Haynes 2004 Sensitive	**ORIGINAL**	sensitiv:.mp. OR diagnos:.mp. OR di.fs.				
	Ritchie 2007	NR	Ovid	69	1.3	
	Leeflang 2006	sensitiv*[Title/Abstract] OR sensitivity and specificity[MeSH Terms] OR diagnos*[Title/Abstract] OR diagnosis[MeSH:noexp] OR diagnostic * [MeSH:noexp] OR diagnosis, differential[MeSH:noexp] OR diagnosis[Subheading:noexp]	PubMed	87		

(Continued)

Whiting 2010	Sensitive$.ti,ab. "sensitivity and specificity"/ Diagnos$.ti,ab. Diagnosis/ Diagnostic$.hw. Diagnosis, Differential/ di.fs.	Ovid	82	3	NNR 36
Noel-Storr 2011	Sensitive$.ti,ab. "sensitivity and specificity"/ Diagnos$.ti,ab. Diagnosis/ Diagnostic$.hw. Diagnosis, Differential/ di.fs.	Ovid	69 (60-77)	0.92 (0.74-1.10)	
Mitchell 2005	1. sensitiv$.mp. 2. diagnos$.mp. 3. di.fs. 4. or/1-3.	Ovid	67	9.1	
Kastner 2009	sensitiv:.mp. OR diagnos:.mp. OR di.fs.	Ovid	88		
Doust 2005	sensitiv:.mp. OR diagnos:.mp. OR di.fs.	Ovid	100		Methodological filter for NPSR
			100	5	Methodological & content filter for NPSR
			88		Methodological filter for TSR
			70	4	Methodological & content filter for NPSR
Haynes 2004	**ORIGINAL**	Specifity.tw.			

(Continued)

Specific	Ritchie 2007	NR	Ovid	21	6.7	
	Whiting 2010	Specificity.ti,ab.	Ovid	43	15	NNR 7
	Noel-Storr 2011	Specificity.ti,ab.	Ovid	14 (9-21)	2.04 (1.22-3.21)	
Haynes 1994 Accurate	**ORIGINAL**	Exp Sensitivity a#d Specificity Or Diagnosis (sh) Or Diagnostic Use (sh) Or Specificity (tw) Or Predicitive (tw) and Value: (tw)				
	Leeflang 2006	‘‘sensitivity and specificity’’[MeSH] OR ‘‘Diagnosis’’[MeSH] OR ‘‘diagnostic use’’[subheading] OR specificity[tw] OR (predictive[tw] AND value[tw])	PubMed	81		
Haynes 1994 Specific	**ORIGINAL**	Exp Sensitivity a#d Specificity OR Predictive (tw) AND Value: (tw)				
	Ritchie 2007	NR	Ovid	33	7.4	
	Leeflang 2006	‘‘sensitivity and specificity’’[MeSH] OR (predictive[tw] AND value[tw])	PubMed	29		
	Whiting 2010	exp "sensitivity and specificity"/ (predictive and value$).ti,ab.	Ovid	56	11	NNR 9
	Noel-Storr 2011	exp "sensitivity and specificity"/ (predictive and value$).ti,ab.	Ovid	51 (42-60)	3.04 (2.36-3.86)	
Haynes 1994 Sensitive	**ORIGINAL**	Exp Sensitivity a#d Specificity or Diagnosis& (px) or Diagnostic Use (sh) or Sensitivity (tw) or Specificity (tw)				
	Ritchie 2007	NR	Ovid	70	1.5	
	Leeflang 2006	"sensitivity and specificity"[MeSH] OR diagnosis[subheading:noexp] OR "diagnostic use"[subheading] OR sensitivity[tw] OR specificity[tw]	PubMed	81		
	Kassai 2006	NR	PubMed	95		

(Continued)

	Whiting 2010	exp "sensitivity and specificity"/ di.xs. Du.fs. Sensitivity.ti,ab. Specificity.ti,ab.	Ovid	87	2	NNR 45
	Vincent 2003	1 exp 'sensitivity and specificity'/ 2 sensitivity.tw. 3 di.fs. 4 du.fs. 5 specificity.tw. 6 or/1-5	NR	96		
	Deville 2000	Sensitivity and specificity (exploded) (sh) Diagnosis& (sh) Diagnostic use (sh) Sensitivity (tw) Specificity (tw)	NR	73 (63-8)		Specificity=94.3 (93.3-95.2); DOR=45
	Noel-Storr 2011	exp "sensitivity and specificity"/ di.xs. Du.fs. Sensitivity.ti,ab. Specificity.ti,ab.	Ovid	91 (84-95)	0.98 (0.80-1.17)	
	Mitchell 2005	1. exp "Sensitivity and Specificity"/ 2. di.xs. 3. du.fs. 4. sensitivity.tw. 5. specificity.tw. 6. or/1-5	Ovid	80	5.3	
Falck-Ytter 2004	**ORIGINAL**	sensitiv:.tw. or exp "sensitivity and specificity"/ or diagnos:.tw,ot,hw,rw. or (di or du).fs.				
	Ritchie 2007	NR	Ovid	74	1.3	
	Whiting 2010	Sensitive:.tw. exp "sensitivity and specificity"/ Diagnos:.tw,ot,hw,rw.	Ovid	85 (80-93)	3 (1-19)	NNR 36 (5-106)

(Continued)

		(di or du).fs.				
	Noel-Storr 2011	Sensitive:.tw. exp "sensitivity and specificity"/ Diagnos:.tw,ot,hw,rw. (di or du).fs.	Ovid	71 (62-79)	1.06 (0.86-1.31)	
Deville 2000 Strategy 1	**ORIGINAL**	sensitivity and specificity (exploded)(sh) specificity (tw)				
	Kassai 2006	NR	PubMed	75.5		
Deville 2000 Strategy 3	**ORIGINAL**	sensitivity and specificity (exploded)(sh) Specificity (tw) False negative (tw) Accuracy (tw)				
	Leeflang 2006	"sensitivity and specificity"[MeSH] OR specificity[tw] OR false negative[tw] OR accuracy [tw]	PubMed	41		
Deville 2000 Strategy 4	**ORIGINAL**	sensitivity and specificity (exploded) (sh) specificity (tw) false negative (tw) accuracy (tw) screening (tw)				
	Ritchie 2007	NR	Ovid	46	4.4	
	Leeflang 2006	‘‘sensitivity and specificity’’[MeSH] OR specificity[tw] OR false negative[tw] OR accuracy[tw] OR screening[tw]	PubMed	46		
	Doust 2005	Sensitivity and specificity [MeSH] Specificity [tw] False negative [tw] Accuracy [tw] Screening [tw]	Ovid	58	9	Methodological & content filter for TSR
				100	9	Methodological & content filter for NPSR
				100		Methodological filter for NPSR
				73		Methodological filter for TSR
	Whiting 2010	exp "sensitivity and specificity"/	Ovid	68	7	NNR 14

(Continued)

		Specificity.ti,ab. False negative.ti,ab. Accuracy.ti,ab. Screening.ti,ab.				
	Vincent 2003	1 exp sensitivity and specificity/ 2 specificit$.tw. 3 false negative$.tw. 4 Accuracy.tw. 5 screening.tw. 6 or/1-5	NR	75		Authors say they tested the Deville specific strategy; however they have listed Deville sensitive strategy in the appendix.
	Noel-Storr 2011	exp "sensitivity and specificity"/ Specificity.ti,ab. False negative.ti,ab. Accuracy.ti,ab. Screening.ti,ab.	Ovid	55 (46-64)	2.20 (1.70-2.77)	
	Mitchell 2005	1. exp "Sensitivity and Specificity"/ 2. specificity.tw. 3. false negative.tw. 4. accuracy.tw. 5. screening.tw. 4. or/1-5	Ovid	49	16.7	
Deville 2002 Extended	**ORIGINAL**	((((((((((("sensitivity and specificity"[All Fields] OR "sensitivity and specificity/standards"[All Fields]) OR "specificity"[All Fields]) OR "screening"[All Fields]) OR "false positive"[All Fields]) OR "false negative"[All Fields]) OR "accuracy"[All Fields]) OR ((((("predictive value"[All Fields] OR "predictive value of tests"[All Fields]) OR "predictive value of tests/standards"[All Fields]) OR "predictive values"[All Fields]) OR "predictive values of tests"[All Fields])) OR (("reference value"[All Fields] OR "reference values"[All Fields]) OR"reference values/ standards"[All Fields])) OR ((((((((((("roc"[All Fields] OR "roc analyses"[All Fields]) OR "roc analysis"[All Fields]) OR "roc and"[All Fields]) OR "roc area"[All Fields]) OR "roc auc"[All Fields]) OR "roc characteristics"[All Fields]) OR				

(Continued)

	"roc curve"[All Fields]) OR "roc curve method"[All Fields]) OR "roc curves"[All Fields]) OR "roc estimated"[All Fields]) OR "roc evaluation"[All Fields])) OR "likelihood ratio"[All Fields]) AND notpubref [sb]) AND "human"[MeSH Terms])				
Ritchie 2007	NR	Ovid	52	3.9	
Doust 2005	((((((((((("sensitivity and specificity"[All Fields] OR "sensitivity and specificity/standards"[All Fields]) OR "specificity"[All Fields]) OR "screening"[All Fields]) OR "false positive"[All Fields]) OR "false negative"[All Fields]) OR "accuracy"[All Fields]) OR ((((("predictive value"[All Fields] OR "predictive value of tests"[All Fields]) OR "predictive value of tests/standards"[All Fields]) OR "predictive values"[All Fields]) OR "predictive values of tests"[All Fields])) OR (("reference value"[All Fields] OR "reference values"[All Fields]) OR"reference values/ standards"[All Fields])) OR ((((((((((("roc"[All Fields] OR "roc analyses"[All Fields]) OR "roc analysis"[All Fields]) OR "roc and"[All Fields]) OR "roc area"[All Fields]) OR "roc auc"[All Fields]) OR "roc characteristics"[All Fields]) OR "roc curve"[All Fields]) OR "roc curve method"[All Fields]) OR "roc curves"[All Fields]) OR "roc estimated"[All Fields]) OR "roc evaluation"[All Fields])) OR "likelihood ratio"[All Fields]) AND notpubref [sb]) AND "human"[MeSH Terms])	WebSpirs	58	8	Methodological & content filter for TSR
			100	6	Methodological & content filter for NPSR
			100		Methodological filter for NPSR
			76		Methodological filter for TSR
Whiting 2010	“sensitivity and specificity”.mp. “sensitivity and specificity”/st Specificity.mp. Screening.mp. False positive.mp. False negative.mp. Accuracy.mp. Predictive value.mp. Predictive values.mp. Reference value.mp. Reference values.mp. Roc.mp. Likelihood ratio.mp.	Ovid	71	7	NNR 15

(Continued)

		Humans/			
	Noel-Storr 2011	"sensitivity and specificity".mp. "sensitivity and specificity"/st Specificity.mp. Screening.mp. False positive.mp. False negative.mp. Accuracy.mp. Predictive value.mp. Predictive values.mp. Reference value.mp. Reference values.mp. Roc.mp. Likelihood ratio.mp. Humans/	Ovid	60 (51-69)	1.99 (1.57-2.47)
Deville 2002a Accurate	**ORIGINAL**	1. sensitivity and specificity[Mesh; exploded] 2. mass screening [Mesh; exploded] 3. reference values [Mesh] 4. false positive reactions [Mesh] 5. false negative reactions [Mesh] 6. specificit$.tw 7. screening.tw 8. false positive$.tw 9. false negative$.tw 10. accuracy.tw 11. predictive value$.tw 12. reference value$.tw 13. roc$.tw 14. likelihood ratio$.tw or/1-14			
	Leeflang 2006	''Sensitivity and Specificity''[MeSH] OR ''mass screening''[MeSH] OR ''Reference values''[MeSH] OR specificit*[tw] OR screening[tw] OR false positive*[tw] OR false negative*[tw] OR accuracy[tw]	PubMed	51	

(Continued)

		OR predictive value*[tw] OR reference value*[tw] OR roc*[tw] OR likelihood ratio*[tw]			
Vincent 2003 Strategy A	**ORIGINAL**	1. exp 'sensitivity and specificity'/ 2. (sensitivity or specificity or accuracy).tw. 3. ((predictive adj3 value$) or (roc adj curve $)).tw. 4. ((false adj positiv$) or (false negativ$)).tw. 5. (observer adj variation$) or (likelihood adj3 ratio$)).tw. 6. likelihood function/ 7. exp mass screening/ 8. diagnosis, differential/ or exp Diagnostic errors/ 9. di.xs or du.fs 10. or/1-9			
	Ritchie 2007	NR	Ovid	87	3.3
	Mitchell 2005	1. exp “Sensitivity and Specificity”/ 2. (sensitivity or specificity or accuracy).tw. 3. ((predictive adj3 value$) or (roc adj curve $)).tw. 4. ((false adj positiv$) or (false negativ$)).tw. 5. (observer adj variation$) or (likelihood adj3 ratio$)).tw. 6. Likelihood Function/ 7. exp Mass Screening/ 8. Diagnosis, Differential/ or exp Diagnostic Errors/ 9. di.xs or du.fs 10. or/1-9	Ovid	81	5.5
Vincent 2003 Strategy C	**ORIGINAL**	1. exp ‘sensitivity and specificity’/ 2. sensitivity.tw. or specificity.tw. 3. (predictive adj3 value$).tw. 4. exp Diagnostic errors/ 5. ((false adj positive$) or (false adj negative$)).tw.			

(Continued)

	6. (observer adj variation$).tw. 7. (roc adj curve$).tw. 8. (likelihood adj3 ratio$).tw. 9. likelihood function/ 10. exp *venous thrombosis/di, ra, ri, us 11. exp *thrombophlebitis/di, ra, ri, us 12. or/1-11				
Leeflang 2006	‘‘sensitivity and specificity’’[MeSH] OR sensitivity[tw] OR specificity[tw] OR predictive value*[tw] OR false positiv*[tw] OR false negativ*[tw] OR observer variation*[tw] OR roc curve*[tw] OR likelihood ratio*[tw] OR ‘‘Likelihood Functions’’[MeSH]	PubMed	44		
Whiting 2010	exp "sensitivity and specificity"/ Sensitivity.tw. Specificity.tw. (predictive adj3 value$).tw. Exp diagnostic errors/ (false adj positiv$).tw. (false adj negativ$).tw. (observer adj variation$).tw. (roc adj curve$).tw. (likelihood adj3 ratio$).tw. Likelihood functions/	Ovid	67	9	NNR 12
Noel-Storr 2011	exp "sensitivity and specificity"/ Sensitivity.tw. Specificity.tw. (predictive adj3 value$).tw. Exp diagnostic errors/ (false adj positiv$).tw. (false adj negativ$).tw. (observer adj variation$).tw. (roc adj curve$).tw. (likelihood adj3 ratio$).tw. Likelihood functions/	Ovid	54 (45-63)	2.30 (1.79-2.89)	

(Continued)

van der Weijden 1997 Sensitive	**ORIGINAL**	*MeSH terms* explode DIAGNOSIS/all.s explode SENSITIVITY-AND-SPECIFICITY REFERENCE-VALUES/all.s FALSE-NEGATIVE-REACTIONS/all.s FALSE-POSITIVE-REACTIONS/all.s explode MASS-SCREENING/all.s *Freetext terms* diagnos* sensitivity or specificity predictive value* reference value* ROC* Likelihood ratio* monitoring				
Leeflang 2006	"Diagnosis"[MeSH] OR "sensitivity and specificity"[MeSH] OR "Reference values"[MeSH] OR "False Positive Reactions"[MeSH] OR "False Negative Reactions"[MeSH] OR "Mass Screening"[MeSH] OR diagnos* OR sensitvity OR specificity OR predictive value* OR reference value* OR ROC* OR likelihood ratio* OR monitoring	PubMed	92			
Doust 2005	Diagnosis [subheading] Sensitivity and Specificity [MeSH] Sensitivity [tw] Specificity [tw] Diagnosis differential [MeSH] Reference values [MeSH] False negative reactions [MeSH] False positive reactions [MeSH] Mass screening [MeSH] diagnos* [tw] predictive value [tw] reference value* [tw] ROC* [tw]	CD-ROM Ovid	Error noted in strategy – original does not include Diagnosis differential [MeSH] and Doust has omitted to add likelihood ratio* and monitoring textwords			
			73	4	Methodological & content filter for TSR	
			100	4	Methodological & content filter for NPSR	
			91		Methodological filter for TSR	
			100		Methodological	

(Continued)

					filter for NPSR
Whiting 2010	Exp diagnosis/ exp "sensitivity and specificity"/ Reference values/ False negative reactions/ False positive reactions/ Exp Mass screening/ Diagnos$.ti,ab. Sensitivity.ti,ab. Specificity.ti,ab. Predictive value$.ti,ab. Reference value$.ti,ab. Roc$.ti,ab. Likelihood ratio$.ti,ab. Monitoring.ti,ab.	Ovid	87	2	NNR 50
Noel-Storr 2011	Exp diagnosis/ exp "sensitivity and specificity"/ Reference values/ False negative reactions/ False positive reactions/ Exp Mass screening/ Diagnos$.ti,ab. Sensitivity.ti,ab. Specificity.ti,ab. Predictive value$.ti,ab. Reference value$.ti,ab. Roc$.ti,ab. Likelihood ratio$.ti,ab. Monitoring.ti,ab.	Ovid	93 (87-97)	0.98 (0.80-1.17)	
Mitchell 2005	exp Diagnosis/ exp "Sensitivity and Specificity"/ Reference Values/ False Negative Reactions/	Ovid	96	5.6	

(Continued)

		False Positive Reactions/ exp Mass Screening/ diagnos$.ti,ab. sensitivity.ti,ab. specificity.ti,ab. predictive value$.ti,ab. reference value$.ti,ab. roc$.ti,ab. likelihood ratio$.ti,ab. monitoring.ti,ab.				
CASP 2002 $	**ORIGINAL**	sensitivity-specificity (s) sensitivity (t) di.fs. ri.fs du.fs specificity (t)				
	Ritchie 2007	NR	Ovid	73	1.2	
	Kassai 2006	NR	PubMed	95		
	Whiting 2010	"sensitivity and specificity"/ Sensitivity.ti,ab. di.fs. Ri.fs. Du.fs. Specificity.ti,ab.	Ovid	83 (78-95)	3 (1-24)	NNR 29 (4-89)
	Vincent 2003	1 exp 'sensitivity and specificity/ 2 sensitivity.tw. 3 di.xs. 4 du.fs. 5 specificity.tw. 6 or/1-5	NR	100		
	Noel-Storr 2011	"sensitivity and specificity"/ Sensitivity.ti,ab. di.fs. Ri.fs. Du.fs. Specificity.ti,ab.	Ovid	67 (58-75)	0.97 (0.77-1.19)	

(Continued)

InterTASC 2011 Aberdeen$	**ORIGINAL**	*MeSH* Exp sensitivity and specificity/ False positive reactions/ False negative reactions/ Du.fs *Text words .tw* Sensitivity Distinguish$ Differentiat$ enhancement Predictive adj4 value$ Identif$ Detect$ Diagnos$ Compar$				
	Ritchie 2007	NR	Ovid	69	1.2	
	Whiting 2010	exp "sensitivity and specificity"/ False positive reactions/ False negative reactions/ Du.fs. Sensitivity.tw. (Predictive adj4 value$).tw. Distinguish$.tw. Differential$.tw. Enhancement.tw. Identif$.tw. Detect$.tw. Diagnos$.tw. Compare$.t	Ovid	86 (81-94)	3 (1-19)	NNR 35 (5-97)
	Noel-Storr 2011	exp "sensitivity and specificity"/ False positive reactions/ False negative reactions/ Du.fs. Sensitivity.tw. (Predictive adj4 value$).tw. Distinguish$.tw. Differential$.tw. Enhancement.tw. Identif$.tw. Detect$.tw. Diagnos$.tw. Compare$.t	Ovid	87 (80-92)	0.95 (0.78-1.14)	
Inter-TASC 2011 Southampton A$ Unclear how terms combined	**ORIGINAL**	*MeSH* Exp sensitivity and specificity/ False positive reactions/ False negative reactions Exp diagnosis/ Reference-values Exp mass screening/ *Text words* Diagnos*				

(Continued)

	Sensitivity Specificity 'sensitivity and specificity' predictive value* Reference value* Roc Roc in AD (NOT) Likelihood ratio* Monitoring				
Ritchie 2007	NR	Ovid	71	1.0	
Whiting 2010	Exp diagnosis/ exp "sensitivity and specificity"/ Reference values/ False negative reactions/ False positive reactions/ Exp mass screening/ Diagnos$.mp. Sensitivity.mp. Specificity.mp. Predictive value$.mp. Reference value$.mp. Roc.mp. NOT roc.in. Likelihood ratio$.mp. Monitoring.mp.	Ovid	86	2	NNR 51
Noel-Storr 2011	Exp diagnosis/ exp "sensitivity and specificity"/ Reference values/ False negative reactions/ False positive reactions/ Exp mass screening/ Diagnos$.mp. Sensitivity.mp. Specificity.mp. Predictive value$.mp. Reference value$.mp. Roc.mp. NOT roc.in. Likelihood ratio$.mp.	Ovid	93 (87-97)	0.96 (0.80-1.15)	

(Continued)

		Monitoring.mp.				
Inter-TASC 2011 Southampton B$ Unclear how terms combined	**ORIGINAL**	*MeSH* Exp sensitivity and specificity/ *Text words* Specificity False negative Accuracy screening				
	Ritchie 2007	NR	Ovid	45	4.6	
	Whiting 2010	exp "sensitivity and specificity"/ Specificity.mp. False negative.mp. Accuracy.mp. Screening.mp.	Ovid	69	7	NNR 14
	Noel-Storr 2011	exp "sensitivity and specificity"/ Specificity.mp. False negative.mp. Accuracy.mp. Screening.mp.	Ovid	55 (46-64)	2.09 (1.63-2.63)	
InterTASC 2011 Southampton C$ Unclear how terms combined	**ORIGINAL**	*MeSH* Exp sensitivity and specificity/ *Text words ti,ab,mesh* Predictive and value				
	Ritchie 2007	NR	Ovid	31	8.5	
	Whiting 2010	exp "sensitivity and specificity"/ (Predictive and value$).ti,ab,sh.	Ovid	56	11	NNR 9
	Noel-Storr 2011	exp "sensitivity and specificity"/ (Predictive and value$).ti,ab,sh.	Ovid	51 (42-60)	3.04 (2.36-3.86)	
Inter-TASC 2011 Southampton D$ Unclear how terms combined	**ORIGINAL**	*MeSH* Exp sensitivity and specificity/ Exp diagnosis/ Exp pathology/ *Text words* Sensitivity Specificity				
	Ritchie 2007	NR	Ovid	66	1.1	

(Continued)

	Whiting 2010	exp "sensitivity and specificity"/ Exp diagnosis/ Exp pathology/ Sensitivity.mp. Specificity.mp.	Ovid	84	2	NNR 48
	Noel-Storr 2011	exp "sensitivity and specificity"/ Exp diagnosis/ Exp pathology/ Sensitivity.mp. Specificity.mp.	Ovid	89 (82-94)	1.13 (0.93-1.35)	
Inter-TASC 2011 Southampton E[$] Unclear how terms combined	**ORIGINAL**	*MeSH* Exp Diagnosis/ Exp sensitivity and specificity False positive reactions/ False negative reactions/ *Text words ti,ab* Diagnos$ ti,ab hw Specificit$ Sensitivit$ Predictive value$ Roc Sroc Receiver operat$ charactristic$ Receiver oprat$ adj2 curve False positiv$ False negative$ accuracy				
	Ritchie 2007	NR	Ovid	71	1.0	
	Whiting 2010	Exp diagnosis/ exp "sensitivity and specificity"/ False positive reactions/ False negative reactions/ Diagnos$.ti,ab,hw. Specificit$.ti,ab. Sensitivit$.ti,ab. Predictive value$.ti,ab. Roc.ti,ab. Sroc.ti,ab. Receiver operat$ characteristic$.ti,ab. (Receiver operat$ adj2 curve).ti,ab	Ovid	87	2	NNR 50

(Continued)

		False positiv$.ti,ab. False negative$.ti,ab. Accuracy.ti,ab.			
	Noel-Storr 2011	Exp diagnosis/ exp "sensitivity and specificity"/ False positive reactions/ False negative reactions/ Diagnos$.ti,ab,hw. Specificit$.ti,ab. Sensitivit$.ti,ab. Predictive value$.ti,ab. Roc.ti,ab. Sroc.ti,ab. Receiver operat$ characteristic$.ti,ab. (Receiver operat$ adj2 curve).ti,ab False positiv$.ti,ab. False negative$.ti,ab. Accuracy.ti,ab.	Ovid	92 (86-96)	0.98 (0.81-1.17)
InterTASC 2011 CRD A Unclear how terms combined	**ORIGINAL**	*MeSH* Exp sensitivity and specificity/ all sub-headings Exp diagnostic errors/ all subheadings *Text Words .ti,ab* Predictive value* Reproducibility Logistic regression Ability near predict* Logistic model* Sroc Roc Positive rate Positive rates Likelihood ratio* Negative rate Negative rates Receiver operating characteristic Correlation Correlated Test or tests near accuracy Curve Curves Test outcome Pretest probabilities Posttest probabilities Roc-curve.mp Logistic-models.mp			

(Continued)

	Likelihood-functions.mp diagnosis				
Ritchie 2007	NR	Ovid	53	2.2	
Whiting 2010	exp "sensitivity and specificity"/ Exp diagnostic errors/ Predictive value$.ti,ab. Reproducibility.ti,ab. Logistic regression.ti,ab. (ability adj5 predict$).ti,ab. Logistic model$.ti,ab. Sroc.ti,ab. Roc.ti,ab. Positive rate.ti,ab. Positive rates.ti,ab. Likelihood ratio$.ti,ab. Negative rate.ti,ab. Negative rates.ti,ab. Receiver operating characteristic.ti,ab. correlation.ti,ab. correlated.ti,ab. ((test or tests) adj5 accuracy).ti,ab. curve.ti,ab. curves.ti,ab. Test outcome.ti,ab. Pretest probabilities.ti,ab. Posttest probabilities.ti,ab. Roc curve.mp. Logistic models.mp. Likelihood functions.mp. diagnosis.ti,ab.	Ovid	73	4	NNR 26
Noel-Storr 2011	exp "sensitivity and specificity"/ Exp diagnostic errors/	Ovid	70 (62-78)	1.23 (0.99-1.50)	

(Continued)

		Predictive value$.ti,ab. Reproducibility.ti,ab. Logistic regression.ti,ab. (ability adj5 predict$).ti,ab. Logistic model$.ti,ab. Sroc.ti,ab. Roc.ti,ab. Positive rate.ti,ab. Positive rates.ti,ab. Likelihood ratio$.ti,ab. Negative rate.ti,ab. Negative rates.ti,ab. Receiver operating characteristic.ti,ab. correlation.ti,ab. correlated.ti,ab. ((test or tests) adj5 accuracy).ti,ab. curve.ti,ab. curves.ti,ab. Test outcome.ti,ab. Pretest probabilities.ti,ab. Posttest probabilities.ti,ab. Roc curve.mp. Logistic models.mp. Likelihood functions.mp. diagnosis.ti,ab.
InterTASC 2011 CRD B Unclear how terms combined	**ORIGINAL**	*MeSH* Exp sensitivity and specificity/ Predictive value of tests/ Logistic models/ Roc curve/ Likelihood functions/ Reference standards/ Reference values/ Severity of illness index/ Reproducibility of results/ Observer variation/ Decision making/ *Text words ti,ab* Diagnos* near5 efficac*

(Continued)

	Diagnos* near5 efficien* Diagnos* near5 effective* Diagnos* near5 accura* Diagnos* near5 correct* Diagnos* near5 reliable Diagnos* near5 reliability Diagnos* near5 error* Diagnos* near5 mistake* Diagnos* near5 inaccura* Diagnos* near5 incorrect Diagnos* near5 unreliable Decision making Sensitivity near5 test Sensitivity near5 tests Specificity near5 test Specificity near5 tests Predictive standard* Predictive value* Predictive model* Predictive factor* Roc Reliability near2 standard* Reliability near2 score* Reliability near2 tool* Reliability near2 aid Reliability near2 aids Performance near2 test Performance near2 tests Performance near2 testing Performance near2 standard* Performance near2 score* Performance near2 tool* Performance near2 aid Performance near2 aids Reference value* sroc Receiver operat* characteristic Receiver operat* curve Likelihood ratio*				
Ritchie 2007	NR	Ovid	40	4.1	
Whiting 2010	exp "sensitivity and specificity"/ Predictive value of tests/ Logistic models/ Roc curve/ Likelihood functions/ Reference standards/ Reference values/ Severity of illness index/ Reproducibility of results/ Observer variation/	Ovid	64	7	NNR 15

(Continued)

Decision making/

(Diagnos$ adj5 efficac$).ti,ab.

(Diagnos$ adj5 efficien$).ti,ab.

(Diagnos$ adj5 effective$).ti,ab.

(Diagnos$ adj5 accura$).ti,ab.

(Diagnos$ adj5 correct$).ti,ab.

(Diagnos$ adj5 reliable).ti,ab.

(Diagnos$ adj5 reliability).ti,ab.

(Diagnos$ adj5 error$).ti,ab.

(Diagnos$ adj5 mistake$).ti,ab.

(Diagnos$ adj5 inaccura$).ti,ab.

(Diagnos$ adj5 incorrect).ti,ab.

(Diagnos$ adj5 unreliable).ti,ab.

Decision making.ti,ab.

(sensitivity adj5 test).ti,ab.

(sensitivity adj5 tests).ti,ab.

(specificity adj5 test).ti,ab.

(specificity adj5 tests).ti,ab.

Predictive standard$.ti,ab.

Predictive value$.ti,ab.

Predictive model$.ti,ab.

Predictive factor$.ti,ab.

Roc.ti,ab.

Receiver operat$ characteristic.ti,ab.

Receiver operat$ curve.ti,ab.

Likelihood ratio$.ti,ab.

Likelihood function.ti,ab.

(false adj2 reaction$).ti,ab.

False positive$.ti,ab.

False negative$.ti,ab.

Gold standard$.ti,ab.

Reference test.ti,ab.

Reference tests.ti,ab.

Reference standard$.ti,ab.

(Continued)

Criter$ standard$.ti,ab.

Criter$ bias.ti,ab.

Criter$ test.ti,ab.

Criter$ tests.ti,ab.

Validat$ standard$.ti,ab.

Validat$ test.ti,ab.

Validat$ tests.ti,ab.

Validat$ bias.ti,ab.

Verificat$ bias.ti,ab.

Work?up bias.ti,ab.

Expectation bias.ti,ab.

Indeterminate result$.ti,ab.

(observer adj2 bias) .ti,ab.

(observer adj10 different) .ti,ab.

Observer variat$.ti,ab.

Interrater reliability.ti,ab.

Interater reliability.ti,ab.

Observer reliability.ti,ab.

(intra$ adj4 reliability) .ti,ab.

(accura$ adj2 test).ti,ab.

(accura$ adj2 tests).ti,ab.

(accura$ adj2 testing).ti,ab.

(accura$ adj2 standard$).ti,ab.

(accura$ adj2 score$).ti,ab.

(accura$ adj2 tool$).ti,ab.

(accura$ adj2 aid).ti,ab.

(accura$ adj2 aids).ti,ab.

(reliability adj2 test).ti,ab.

(reliability adj2 tests).ti,ab.

(reliability adj2 testing).ti,ab.

(reliability adj2 standard$).ti,ab.

(reliability adj2 score$).ti,ab.

(reliability adj2 tool$).ti,ab.

(reliability adj2 aid).ti,ab.

(Continued)

	(reliability adj2 aids).ti,ab. (performance adj2 test).ti,ab. (performance adj2 tests).ti,ab. (performance adj2 testing).ti,ab. (performance adj2 standard$).ti,ab. (performance adj2 score$).ti,ab. (performance adj2 tool$).ti,ab. (performance adj2 aid).ti,ab. (performance adj2 aids).ti,ab. Reference value$.ti,ab. Sroc.ti,ab.			
Noel-Storr 2011	exp "sensitivity and specificity"/ Predictive value of tests/ Logistic models/ Roc curve/ Likelihood functions/ Reference standards/ Reference values/ Severity of illness index/ Reproducibility of results/ Observer variation/ Decision making/ (Diagnos$ adj5 efficac$).ti,ab. (Diagnos$ adj5 efficien$).ti,ab. (Diagnos$ adj5 effective$).ti,ab. (Diagnos$ adj5 accura$).ti,ab. (Diagnos$ adj5 correct$).ti,ab. (Diagnos$ adj5 reliable).ti,ab. (Diagnos$ adj5 reliability).ti,ab. (Diagnos$ adj5 error$).ti,ab. (Diagnos$ adj5 mistake$).ti,ab. (Diagnos$ adj5 inaccura$).ti,ab. (Diagnos$ adj5 incorrect).ti,ab. (Diagnos$ adj5 unreliable).ti,ab.	Ovid	67 (58-75)	1.69 (1.40-2.10)

(Continued)

Decision making.ti,ab.

(sensitivity adj5 test).ti,ab.

(sensitivity adj5 tests).ti,ab.

(specificity adj5 test).ti,ab.

(specificity adj5 tests).ti,ab.

Predictive standard$.ti,ab.

Predictive value$.ti,ab.

Predictive model$.ti,ab.

Predictive factor$.ti,ab.

Roc.ti,ab.

Receiver operat$ characteristic.ti,ab.

Receiver operat$ curve.ti,ab.

Likelihood ratio$.ti,ab.

Likelihood function.ti,ab.

(false adj2 reaction$).ti,ab.

False positive$.ti,ab.

False negative$.ti,ab.

Gold standard$.ti,ab.

Reference test.ti,ab.

Reference tests.ti,ab.

Reference standard$.ti,ab.

Criter$ standard$.ti,ab.

Criter$ bias.ti,ab.

Criter$ test.ti,ab.

Criter$ tests.ti,ab.

Validat$ standard$.ti,ab.

Validat$ test.ti,ab.

Validat$ tests.ti,ab.

Validat$ bias.ti,ab.

Verificat$ bias.ti,ab.

Work?up bias.ti,ab.

Expectation bias.ti,ab.

Indeterminate result$.ti,ab.

(observer adj2 bias) .ti,ab.

(Continued)

		(observer adj10 different) .ti,ab. Observer variat$.ti,ab. Interrater reliability.ti,ab. Interater reliability.ti,ab. Observer reliability.ti,ab. (intra$ adj4 reliability) .ti,ab. (accura$ adj2 test).ti,ab. (accura$ adj2 tests).ti,ab. (accura$ adj2 testing).ti,ab. (accura$ adj2 standard$).ti,ab. (accura$ adj2 score$).ti,ab. (accura$ adj2 tool$).ti,ab. (accura$ adj2 aid).ti,ab. (accura$ adj2 aids).ti,ab. (reliability adj2 test).ti,ab. (reliability adj2 tests).ti,ab. (reliability adj2 testing).ti,ab. (reliability adj2 standard$).ti,ab. (reliability adj2 score$).ti,ab. (reliability adj2 tool$).ti,ab. (reliability adj2 aid).ti,ab. (reliability adj2 aids).ti,ab. (performance adj2 test).ti,ab. (performance adj2 tests).ti,ab. (performance adj2 testing).ti,ab. (performance adj2 standard$).ti,ab. (performance adj2 score$).ti,ab. (performance adj2 tool$).ti,ab. (performance adj2 aid).ti,ab. (performance adj2 aids).ti,ab. Reference value$.ti,ab. Sroc.ti,ab.
InterTASC 2011	**ORIGINAL**	*MeSH* Exp Sensitivity and specificity/ False positive reactions/

(Continued)

CRD C Unclear how terms combined	False negative reactions/ Logistic models/ Roc curve/ Likelihood functions/ diagnosis/ Exp diagnostic errors/ Exp diagnostic techniques and procedures/ Exp laboratory techniques and procedures/ *Text words ti,ab* Specificit$ Sensitivit$ False negative$ False positive$ True negative$ True positive$ Positive rate$ Negative rate$ Screening Accuracy Reference value$ Likelihood ratio$ Sroc Srocs Roc Rocs Receiver operat$ curve$ Receiver operat$ character$ Diagnos$ adj3 efficac$ Diagnos$ adj3 efficien$ Diagnos$ adj3 effectiv$ Diagnos$ adj3 accura$ Diagnos$ adj3 correct$ Diagnos$ adj3 reliable Diagnos$ adj3 reliability Diagnos$ adj3 error$ Diagnos$ adj3 mistake$ Diagnos$ adj3 inaccura$ Diagnos$ adj3 incorrect Diagnos$ adj3 unreliable Diagnostic yield.mp Misdiagnos$ Reproductivity.mp Logistical regression.mp Logistical model$ Ability adj2 predict$ Reliable adj3 test Reliable adj3 tests Reliable adj3 testing Reliable adj3 standard Reliability adj3 test Reliability adj3 tests Reliability adj3 testing Reliability adj3 standard Performance adj3 test Performance adj3 tests Performance adj3 testing Performance adj3 standard$

(Continued)

	Predictive adj value$ Predictive adj standard$ Predictive adj model$ Predictive adj factor$ Reference adj test Reference adj tests Reference adj testing Index adj test Index adj tests Index adj testing				
Ritchie 2007	NR	Ovid	69	1.2	
Whiting 2010	exp "sensitivity and specificity"/ False positive reactions/ False negative reactions/ Logistic models/ Roc curve/ Likelihood functions/ Diagnosis/ Exp diagnostic errors/ exp "Diagnostic Techniques and Procedures"/ exp "laboratory techniques and procedures"/ Specificit$.ti,ab. Sensitivity$.ti,ab. False negative$.ti,ab. False positive$.ti,ab. True negative$.ti,ab. True positive$.ti,ab. Positive rate$.ti,ab. Negative rate$.ti,ab. Screening.ti,ab. Accuracy.ti,ab. Reference value$.ti,ab. Likelihood ratio$.ti,ab. Sroc.ti,ab. Srocs.ti,ab.	Ovid	85	2	NNR 46

(Continued)

Roc.ti,ab.

Rocs.ti,ab.

Receiver operat$ curve$.ti,ab.

Receiver operat$ character$.ti,ab.

(Diagnos$ adj3 efficac$).ti,ab.

(Diagnos$ adj3 efficien$).ti,ab.

(Diagnos$ adj3 effectiv$).ti,ab.

(Diagnos$ adj3 accura$).ti,ab.

(Diagnos$ adj3 correct$).ti,ab.

(Diagnos$ adj3 reliable).ti,ab.

(Diagnos$ adj3 reliability).ti,ab.

(Diagnos$ adj3 error$).ti,ab.

(Diagnos$ adj3 mistake$).ti,ab.

(Diagnos$ adj3 inaccura$).ti,ab.

(Diagnos$ adj3 incorrect$).ti,ab.

(Diagnos$ adj3 unreliable).ti,ab.

Diagnostic yield.mp.

Misdiagnos$.ti,ab.

Reproductivity.mp.

Logistical regression.mp.

Logistical model$.ti,ab.

(ability adj2 predict$).ti,ab.

(reliable adj3 test).ti,ab.

(reliable adj3 tests).ti,ab.

(reliable adj3 testing).ti,ab.

(reliable adj3 standard).ti,ab.

(reliability adj3 test).ti,ab.

(reliability adj3 tests).ti,ab.

(reliability adj3 testing).ti,ab.

(reliability adj3 standard).ti,ab.

(performance adj3 test).ti,ab.

(performance adj3 tests).ti,ab.

(performance adj3 testing).ti,ab.

(performance adj3 standard$).ti,ab.

(Continued)

	(Predictive adj value$).ti,ab. (Predictive adj standard$).ti,ab. (Predictive adj model$).ti,ab. (Predictive adj factor$).ti,ab. (Reference adj test).ti,ab. (Reference adj tests).ti,ab. (Reference adj testing).ti,ab. (index adj test).ti,ab. (index adj tests).ti,ab. (index adj testing).ti,ab.			
Noel-Storr 2011	exp "sensitivity and specificity"/ False positive reactions/ False negative reactions/ Logistic models/ Roc curve/ Likelihood functions/ Diagnosis/ Exp diagnostic errors/ exp "Diagnostic Techniques and Procedures"/ exp "laboratory techniques and procedures"/ Specificit$.ti,ab. Sensitivity$.ti,ab. False negative$.ti,ab. False positive$.ti,ab. True negative$.ti,ab. True positive$.ti,ab. Positive rate$.ti,ab. Negative rate$.ti,ab. Screening.ti,ab. Accuracy.ti,ab. Reference value$.ti,ab. Likelihood ratio$.ti,ab.	Ovid	90 (83-94)	1.15 (0.95-1.38)

(Continued)

Sroc.ti,ab.

Srocs.ti,ab.

Roc.ti,ab.

Rocs.ti,ab.

Receiver operat$ curve$.ti,ab.

Receiver operat$ character$.ti,ab.

(Diagnos$ adj3 efficac$).ti,ab.

(Diagnos$ adj3 efficien$).ti,ab.

(Diagnos$ adj3 effectiv$).ti,ab.

(Diagnos$ adj3 accura$).ti,ab.

(Diagnos$ adj3 correct$).ti,ab.

(Diagnos$ adj3 reliable).ti,ab.

(Diagnos$ adj3 reliability).ti,ab.

(Diagnos$ adj3 error$).ti,ab.

(Diagnos$ adj3 mistake$).ti,ab.

(Diagnos$ adj3 inaccura$).ti,ab.

(Diagnos$ adj3 incorrect$).ti,ab.

(Diagnos$ adj3 unreliable).ti,ab.

Diagnostic yield.mp.

Misdiagnos$.ti,ab.

Reproductivity.mp.

Logistical regression.mp.

Logistical model$.ti,ab.

(ability adj2 predict$).ti,ab.

(reliable adj3 test).ti,ab.

(reliable adj3 tests).ti,ab.

(reliable adj3 testing).ti,ab.

(reliable adj3 standard).ti,ab.

(reliability adj3 test).ti,ab.

(reliability adj3 tests).ti,ab.

(reliability adj3 testing).ti,ab.

(reliability adj3 standard).ti,ab.

(performance adj3 test).ti,ab.

(performance adj3 tests).ti,ab.

(Continued)

		(performance adj3 testing).ti,ab. (performance adj3 standard$).ti,ab. (Predictive adj value$).ti,ab. (Predictive adj standard$).ti,ab. (Predictive adj model$).ti,ab. (Predictive adj factor$).ti,ab. (Reference adj test).ti,ab. (Reference adj tests).ti,ab. (Reference adj testing).ti,ab. (index adj test).ti,ab. (index adj tests).ti,ab. (index adj testing).ti,ab.				
InterTASC 2011 HTBS Unclear how terms combined	**ORIGINAL**	*MeSH* Exp Sensitivity and specificity/ Exp Diagnostic errors/ Likelihood functions/ Reproducibility of results/ *Text words .tw* Sensitivit$ Specificit$ Accurac$ Predictive adj2 value$ False$ adj2 positive$ False$ adj2 negative$ False$ adj2 rate$ roc Receiver operat$ adj2 curve$ Receiver operat$ characteristic$ Likelihood$ adj2 ratio$ Likelihood$ adj2 function$				
	Ritchie 2007	NR	Ovid	46	3.7	
	Whiting 2010	exp "sensitivity and specificity"/ Exp diagnostic errors/ Likelihood functions/ Reproducibility of results/ Sensitivity$.tw. Specificit$.tw. Accuracy$.tw. (Predictive adj2 value$).tw. (False$ adj2 positive$).tw.	Ovid	69	8	NNR 12

(Continued)

		(false$ adj2 negative$).tw. (false$ adj2 rate$).tw. Roc.tw. (receiver operat$ adj2 curve$).tw. (receiver operat$ characteristic$).tw (likelihood$ adj2 ratio$).tw. (likelihood$ adj2 function$).tw.			
	Noel-Storr 2011	exp "sensitivity and specificity"/ Exp diagnostic errors/ Likelihood functions/ Reproducibility of results/ Sensitivity$.tw. Specificit$.tw. Accuracy$.tw. (Predictive adj2 value$).tw. (False$ adj2 positive$).tw. (false$ adj2 negative$).tw. (false$ adj2 rate$).tw. Roc.tw. (receiver operat$ adj2 curve$).tw. (receiver operat$ characteristic$).tw (likelihood$ adj2 ratio$).tw. (likelihood$ adj2 function$).tw.	Ovid	56 (47-65)	2.04 (1.60-2.57)
Shipley Miner 2002 $	**ORIGINAL**	1 exp "sensitivity and specificity"/ 2 (sensitivity or specificity).ti.ab. 3 likelihood functions/ 4 exp diagnostic errors/ 5 area under curve/ 6 reproducibility of results/ 7 (predictive adj value$1).ti.ab. 8 (likelihood adj ratio$1).ti.ab. 9 (false adj (negative$1 or positive$1).ti.ab. 10 diagnosis, differential/ 11 random allocations/ 12 random$.ti,ab. 13 ((single or double or triple) adj blind $3).ti,ab. 14 double blind method/ or single blind method/ 15 (randomized controlled trial or controlled clinical trial).pt.			

(Continued)

	16 practice guideline.pt. 17 consensus development conference$.pt. 18 1 or 2 or 8 or 3 19 or/1-17				
Ritchie 2007	NR	Ovid	48	1.8	
Whiting 2010	exp "sensitivity and specificity"/ (sensitivity or specificity).ti,ab. Likelihood functions/ Exp diagnostic errors/ Area under curve/ Reproducibility of results/ (predictive adj value$1).ti,ab. (likelihood adj ratio$1).ti,ab. (false adj (negative$1 or positive$1)).ti,ab. Diagnosis, differential/ Random allocation/ Random$.ti,ab. ((single or double or triple) adj blind$3).ti,ab. Double blind method/ Single blind method/ Randomized controlled trial.pt. Controlled clinical trial.pt. Practice guideline.pt. Consensus development conference$.pt.	Ovid	72	5	NNR 19
Noel-Storr 2011	exp "sensitivity and specificity"/ (sensitivity or specificity).ti,ab. Likelihood functions/ Exp diagnostic errors/ Area under curve/ Reproducibility of results/ (predictive adj value$1).ti,ab.	Ovid	63 (54-72)	1.71 (1.35-2.12)	

(Continued)

		(likelihood adj ratio$1).ti,ab. (false adj (negative$1 or positive$1)).ti,ab. Diagnosis, differential/ Random allocation/ Random$.ti,ab. ((single or double or triple) adj blind$3).ti,ab. Double blind method/ Single blind method/ Randomized controlled trial.pt. Controlled clinical trial.pt. Practice guideline.pt. Consensus development conference$.pt.		
University of Rochester 2002 $	**ORIGINAL**	Unable to access – website no longer valid		
	Vincent 2003	1 exp 'sensitivity and specificity'/ 2 false negative reactions/ or false positive reactions/ 3 (sensitivity or specificity).ti,ab. 4 (predicitive adj value$1).ti,ab. 5 (likelihood adj ratio$10.TI,AB. 6 (false adj (negative$1 or positive$1)).ti,ab. 7 or/1-7	NR	79
North Thames 2002	**ORIGINAL**	Unable to access – website no longer valid		
	Vincent 2003	1 exp 'sensitivity and specificity' 2 exp diagnostic errors 3 mass screening 4 or/1-3	NR	53

Abbreviations used: TSR = Tympanometry systematic review; NPSR = Natriuretic peptides systematic review.

$ Filter no longer available from source cited by evaluation studies.

WHAT'S NEW

Date	Event	Description
30 November 2011	Amended	Updated the protocol and added an author
3 July 2009	Amended	Updated the protocol with other authors and revised text
27 December 2007	Amended	Converted to new review format

HISTORY

Protocol first published: Issue 2, 2006
Review first published: Issue 9, 2013

Date	Event	Description
20 February 2007	New citation required and major changes	Substantive amendment

CONTRIBUTIONS OF AUTHORS

Rebecca Beynon designed the study, ran literature searches, screened literature searches, extracted data, synthesized data and drafted the manuscript. Julie Glanville devised and ran literature searches and drafted the manuscript. Mariska Leeflang designed the study, screened literature searches and edited the manuscript. Ruth Mitchell devised and ran literature searches, screened literature searches and edited the manuscript. Anne Eisinga devised and ran literature searches, and edited the manuscript. Steve McDonald devised and ran literature searches, screened literature searches and edited the manuscript. Penny Whiting edited the manuscript.

DECLARATIONS OF INTEREST

Julie Glanville, together with colleagues from the InterTASC Information Specialist Subgroup, developed the Search Filter Appraisal Checklist that is used in this review for the methodological assessment of the included studies and has published search filters. Julie Glanville, Mariska Leeflang, Ruth Mitchell, Rebecca Beynon and Penny Whiting have published performance evaluations of search filters.

SOURCES OF SUPPORT

Internal sources

- National Institute for Health Research, UK.

  Incentive award for completion of review

External sources

- No sources of support supplied

INDEX TERMS

Medical Subject Headings (MeSH)

*Diagnosis; *Subject Headings; Databases, Bibliographic; Information Storage and Retrieval [*methods] [standards]; MEDLINE; Reference Standards; Review Literature as Topic; Search Engine; Sensitivity and Specificity