HEALTH TECHNOLOGY ASSESSMENT

VOLUME 21 ISSUE 69 NOVEMBER 2017
ISSN 1366-5278

Assessing the performance of methodological search filters to improve the efficiency of evidence information retrieval: five literature reviews and a qualitative study

Carol Lefebvre, Julie Glanville, Sophie Beale, Charles Boachie, Steven Duffy, Cynthia Fraser, Jenny Harbour, Rachael McCool and Lynne Smith

NHS
National Institute for Health Research

DOI 10.3310/hta21690

Assessing the performance of methodological search filters to improve the efficiency of evidence information retrieval: five literature reviews and a qualitative study

Carol Lefebvre,[1,2]* Julie Glanville,[3] Sophie Beale,[3] Charles Boachie,[4] Steven Duffy,[3] Cynthia Fraser,[4] Jenny Harbour,[5] Rachael McCool[3] and Lynne Smith[5]

[1]UK Cochrane Centre, Oxford, UK
[2]Lefebvre Associates Ltd, Oxford, UK
[3]York Health Economics Consortium, York, UK
[4]Health Services Research Unit, University of Aberdeen, Aberdeen, UK
[5]Healthcare Improvement Scotland, Glasgow, UK

*Corresponding author

Declared competing interests of authors: none

Note to reader: it is acknowledged that there has been a regrettable delay between carrying out the project, including the searches, and the publication of this report, because of serious illness of the principal investigator. The searches were carried out in 2010/11.

Published November 2017
DOI: 10.3310/hta21690

This report should be referenced as follows:

Lefebvre C, Glanville J, Beale S, Boachie C, Duffy S, Fraser C, *et al*. Assessing the performance of methodological search filters to improve the efficiency of evidence information retrieval: five literature reviews and a qualitative study. *Health Technol Assess* 2017;**21**(69).

Health Technology Assessment is indexed and abstracted in *Index Medicus*/MEDLINE, *Excerpta Medica*/EMBASE, *Science Citation Index Expanded* (SciSearch®) and *Current Contents®*/ Clinical Medicine.

Health Technology Assessment

HTA MRP–MRC

ISSN 1366-5278 (Print)

ISSN 2046-4924 (Online)

Impact factor: 4.236

Health Technology Assessment is indexed in MEDLINE, CINAHL, EMBASE, The Cochrane Library and the Clarivate Analytics Science Citation Index.

This journal is a member of and subscribes to the principles of the Committee on Publication Ethics (COPE) (www.publicationethics.org/).

Editorial contact: journals.library@nihr.ac.uk

The full HTA archive is freely available to view online at www.journalslibrary.nihr.ac.uk/hta. Print-on-demand copies can be purchased from the report pages of the NIHR Journals Library website: www.journalslibrary.nihr.ac.uk

Criteria for inclusion in the *Health Technology Assessment* journal

Reports are published in *Health Technology Assessment* (HTA) if (1) they have resulted from work for the HTA programme or, commissioned/managed through the Methodology research programme (MRP), and (2) they are of a sufficiently high scientific quality as assessed by the reviewers and editors.

Reviews in *Health Technology Assessment* are termed 'systematic' when the account of the search appraisal and synthesis methods (to minimise biases and random errors) would, in theory, permit the replication of the review by others.

HTA programme

The HTA programme, part of the National Institute for Health Research (NIHR), was set up in 1993. It produces high-quality research information on the effectiveness, costs and broader impact of health technologies for those who use, manage and provide care in the NHS. 'Health technologies' are broadly defined as all interventions used to promote health, prevent and treat disease, and improve rehabilitation and long-term care.

The journal is indexed in NHS Evidence via its abstracts included in MEDLINE and its Technology Assessment Reports inform National Institute for Health and Care Excellence (NICE) guidance. HTA research is also an important source of evidence for National Screening Committee (NSC) policy decisions.

For more information about the HTA programme please visit the website: http://www.nets.nihr.ac.uk/programmes/hta

This report

This issue of the Health Technology Assessment journal series contains a project commissioned/managed by the Methodology research programme (MRP). The Medical Research Council (MRC) is working with NIHR to deliver the single joint health strategy and the MRP was launched in 2008 as part of the delivery model. MRC is lead funding partner for MRP and part of this programme is the joint MRC–NIHR funding panel 'The Methodology Research Programme Panel'.

To strengthen the evidence base for health research, the MRP oversees and implements the evolving strategy for high-quality methodological research. In addition to the MRC and NIHR funding partners, the MRP takes into account the needs of other stakeholders including the devolved administrations, industry R&D, and regulatory/advisory agencies and other public bodies. The MRP funds investigator-led and needs-led research proposals from across the UK. In addition to the standard MRC and RCUK terms and conditions, projects commissioned/managed by the MRP are expected to provide a detailed report on the research findings and may publish the findings in the HTA journal, if supported by NIHR funds.

The authors have been wholly responsible for all data collection, analysis and interpretation, and for writing up their work. The HTA editors and publisher have tried to ensure the accuracy of the authors' report and would like to thank the reviewers for their constructive comments on the draft document. However, they do not accept liability for damages or losses arising from material published in this report.

This report presents independent research funded under a MRC–NIHR partnership. The views and opinions expressed by authors in this publication are those of the authors and do not necessarily reflect those of the NHS, the NIHR, the MRC, NETSCC, the HTA programme or the Department of Health. If there are verbatim quotations included in this publication the views and opinions expressed by the interviewees are those of the interviewees and do not necessarily reflect those of the authors, those of the NHS, the NIHR, the MRC, NETSCC, the HTA programme or the Department of Health.

Published by the NIHR Journals Library (www.journalslibrary.nihr.ac.uk), produced by Prepress Projects Ltd, Perth, Scotland (www.prepress-projects.co.uk).

Abstract

Assessing the performance of methodological search filters to improve the efficiency of evidence information retrieval: five literature reviews and a qualitative study

Carol Lefebvre,[1,2]* Julie Glanville,[3] Sophie Beale,[3] Charles Boachie,[4] Steven Duffy,[3] Cynthia Fraser,[4] Jenny Harbour,[5] Rachael McCool[3] and Lynne Smith[5]

[1]UK Cochrane Centre, Oxford, UK
[2]Lefebvre Associates Ltd, Oxford, UK
[3]York Health Economics Consortium, York, UK
[4]Health Services Research Unit, University of Aberdeen, Aberdeen, UK
[5]Healthcare Improvement Scotland, Glasgow, UK

*Corresponding author Carol@LefebvreAssociates.org

Background: Effective study identification is essential for conducting health research, developing clinical guidance and health policy and supporting health-care decision-making. Methodological search filters (combinations of search terms to capture a specific study design) can assist in searching to achieve this.

Objectives: This project investigated the methods used to assess the performance of methodological search filters, the information that searchers require when choosing search filters and how that information could be better provided.

Methods: Five literature reviews were undertaken in 2010/11: search filter development and testing; comparison of search filters; decision-making in choosing search filters; diagnostic test accuracy (DTA) study methods; and decision-making in choosing diagnostic tests. We conducted interviews and a questionnaire with experienced searchers to learn what information assists in the choice of search filters and how filters are used. These investigations informed the development of various approaches to gathering and reporting search filter performance data. We acknowledge that there has been a regrettable delay between carrying out the project, including the searches, and the publication of this report, because of serious illness of the principal investigator.

Results: The development of filters most frequently involved using a reference standard derived from hand-searching journals. Most filters were validated internally only. Reporting of methods was generally poor. Sensitivity, precision and specificity were the most commonly reported performance measures and were presented in tables. Aspects of DTA study methods are applicable to search filters, particularly in the development of the reference standard. There is limited evidence on how clinicians choose between diagnostic tests. No published literature was found on how searchers select filters. Interviewing and questioning searchers via a questionnaire found that filters were not appropriate for all tasks but were predominantly used to reduce large numbers of retrieved records and to introduce focus. The Inter Technology Appraisal Support Collaboration (InterTASC) Information Specialists' Sub-Group (ISSG) Search Filters Resource was most frequently mentioned by both groups as the resource consulted to select a filter. Randomised controlled trial (RCT) and systematic review filters, in particular the Cochrane RCT and the McMaster Hedges filters, were most frequently mentioned. The majority indicated that they used different filters depending on the requirement for sensitivity or precision. Over half of the respondents used the filters

available in databases. Interviewees used various approaches when using and adapting search filters. Respondents suggested that the main factors that would make choosing a filter easier were the availability of critical appraisals and more detailed performance information. Provenance and having the filter available in a central storage location were also important.

Limitations: The questionnaire could have been shorter and could have included more multiple choice questions, and the reviews of filter performance focused on only four study designs.

Conclusions: Search filter studies should use a representative reference standard and explicitly report methods and results. Performance measures should be presented systematically and clearly. Searchers find filters useful in certain circumstances but expressed a need for more user-friendly performance information to aid filter choice. We suggest approaches to use, adapt and report search filter performance. Future work could include research around search filters and performance measures for study designs not addressed here, exploration of alternative methods of displaying performance results and numerical synthesis of performance comparison results.

Funding: The National Institute for Health Research (NIHR) Health Technology Assessment programme and Medical Research Council–NIHR Methodology Research Programme (grant number G0901496).

Contents

List of tables

List of figures

List of boxes

Glossary

Accuracy The number of records correctly retrieved (because they are relevant) plus the number correctly not retrieved (because they are not relevant) as a proportion of all records in the database. Often expressed as a percentage.

Area under the curve Calculation of the area under the receiver operating characteristic curve provides the overall value of diagnostic test accuracy.

Article read ratio The number of articles (or records) retrieved by a search filter that need to be read to identify one relevant record. This is calculated as 1/precision and is equivalent to the number needed to read.

Diagnostic odds ratio The odds of being truly relevant among the relevant divided by the odds of being assessed as relevant among the irrelevant.

External standard A reference standard used to validate a search filter that is different from the one from which the filter has been derived.

Fallout 1 – specificity value.

Gold standard A collection of records that meet specific criteria for relevance. The criteria for relevance will vary. Performance measures for search filters measure how well the filters retrieve records from the gold standard. Also known as a reference set or standard. When a search filter is developed and its performance is measured on the same gold standard, this standard is described as an internal standard. When a filter is developed and measured using a different gold standard, this standard is described as an external standard.

Hand-searching Assessment of the full texts of publications such as journals to identify relevant records meeting reference set or gold standard inclusion criteria. Hand-searching typically involves the examination of documents from cover to cover for a specified publication time span (in the case of journals).

Hedges An alternative name for search filters.

Internal standard A reference standard that is used to derive and validate a search filter.

Irrelevant records These records may be retrieved by the search filter but do not meet the criteria for inclusion in the reference set/gold standard.

Methodological search filter A search filter designed to retrieve a specific research method.

Multiple technology appraisal An appraisal of the clinical effectiveness and cost-effectiveness of, typically, more than one technology that is undertaken by an independent academic centre commissioned by the National Institute for Health and Care Excellence.

Number needed to read The number of records retrieved by a search filter that need to be read to identify one relevant record. This is calculated as 1/precision.

Number of records retrieved The total number of records retrieved by a search filter.

Precision The number of reference set or gold standard (i.e. relevant) records retrieved by a search filter as a proportion of the total number of records (relevant and irrelevant) retrieved. Often expressed as a percentage.

Prevalence The number of relevant records in the reference set retrieved as a proportion of the total number of records in a database. Often expressed as a percentage.

Recall The number of relevant records in the reference set or gold standard that are retrieved by a search filter as a proportion of the total number of records in the reference set or gold standard. Often expressed as a percentage and also known as sensitivity.

Receiver operating characteristic A receiver operating characteristic curve represents the relationship between the 'true-positive fraction' (sensitivity) and the 'false-positive fraction' (1 – specificity).

Reduction in number needed to read/screen The reduction in the number of retrieved records when a filter is applied, expressed as a percentage of the number retrieved before its application.

Reference set/standard See *Gold standard*.

Reference standard spectrum bias The variation in the sensitivity and/or specificity of a diagnostic test when applied to an unrepresentative sample.

Relative recall gold standard Included studies from a specific review (or other source) that can be used as a test set to test the sensitivity of a search filter.

Relevant records Records from the reference set/gold standard.

Results set The collection of records retrieved by hand-searching or by a search strategy, filter or combination of both (depending on the context). The results set contains relevant and irrelevant records.

Retrieval gain The absolute or percentage variation in the number of records retrieved by the search filter.

Search filter A combination of search terms to identify specific topics (such as breast cancer) or study designs (such as randomised controlled trials) or other issues such as age, gender or geographical area.

Search filter performance A measure of how well a search filter performs in identifying relevant studies or not retrieving irrelevant studies. Measures include accuracy, number needed to read, precision, sensitivity and specificity.

Search question The research topic that the search strategy is seeking to capture. The search question may be more or less specific than the search strategy depending on how much of the search question can be captured by search terms and how many concepts are included in the search strategy.

Sensitivity The number of relevant records in the reference set/gold standard that are retrieved by a search filter as a proportion of the total number of records in the reference set/gold standard. Often expressed as a percentage and also known as recall.

Single technology appraisal A critical appraisal of a manufacturer's assessment of the clinical effectiveness and cost-effectiveness of a single technology. Undertaken by independent academic centres commissioned by the National Institute for Health and Care Excellence.

Specificity The number of irrelevant records correctly not retrieved as a proportion of all irrelevant records in the resource. Often expressed as a percentage.

Study design The methods used within a research study, for example a randomised controlled study design.

Subject search A search strategy containing terms designed to capture a specific topic such as an intervention, a disease, an outcome or a population group. Subject searches may combine several concepts.

Validation (external) See *External standard*.

Validation (internal) See *Internal standard*.

List of abbreviations

AHRQ	Agency for Healthcare Research and Quality
ASSIA	Applied Social Sciences Index and Abstracts
AUC	area under the curve
CADTH	Canadian Agency for Drugs and Technologies in Health
CD-ROM	compact disc, read-only memory
CDSR	Cochrane Database of Systematic Reviews
CENTRAL	Cochrane Central Register of Controlled Trials
CINAHL	Cumulative Index to Nursing and Allied Health Literature
CRD	Centre for Reviews and Dissemination
DARE	Database of Abstracts of Reviews of Effects
DOR	diagnostic odds ratio
DTA	diagnostic test accuracy
EAHIL	European Association for Health Information and Libraries
EBLIP	Evidence Based Library and Information Practice
ERG	Evidence Review Group
EUnetHTA	European network for Health Technology Assessment
FDA	Food and Drug Administration
HEED	Health Economic Evaluations Database
HTA	Health Technology Assessment
HTAi	Health Technology Assessment international
InterTASC	Inter Technology Appraisal Support Collaboration
IRMG	Information Retrieval Methods Group
ISSG	Information Specialists' Sub-Group
LILACS	Latin American and Caribbean Health Sciences Literature
LR	likelihood ratio
LR+	positive likelihood ratio
LR–	negative likelihood ratio
MeSH	medical subject heading
NCC	National Collaborating Centre
NHS EED	NHS Economic Evaluation Database
NICE	National Institute for Health and Care Excellence
NLM	National Library of Medicine
NNR	number needed to read
NPV	negative predictive value
PPV	positive predictive value
QUADAS	Quality Assessment of Diagnostic Accuracy Studies
RCT	randomised controlled trial
ROC	receiver operating characteristic
RR	relative recall
RSS	really simple syndication
SIGN	Scottish Intercollegiate Guidelines Network
SROC	summary receiver operating characteristic
STARD	Standards for the Reporting of Diagnostic Accuracy Studies
TSC	Trials Search Co-ordinator
YHEC	York Health Economics Consortium

Plain English summary

Effective identification of research studies is essential for developing clinical guidance and health policy, conducting health research and supporting health-care decision-making. Methodological search filters (combinations of search terms to identify studies of a specific design) can help to find relevant studies when searching literature databases. This project investigated issues around the creation and performance of methodological search filters and how best to assist searchers in choosing search filters. We conducted five literature reviews in 2010/11, interviewed searchers about their use of search filters and circulated a questionnaire to a larger group of searchers. The findings were used to suggest how best to collect and report data on search filter performance.

We found that studies that created search filters reported sensitivity (the proportion of relevant articles retrieved), precision (the proportion of articles retrieved that are relevant) and specificity (the proportion of non-relevant articles not retrieved) most often. However, it was sometimes difficult to judge the quality of the study design because the authors did not provide an adequate description of how they had conducted their study. In addition, several studies did not use the best methods available; for example, they tested the filter on database records that had been used to create the filter. More detailed reporting and a clearer presentation of the results with graphs would make it easier to judge the reliability of the results.

The majority of searchers who were interviewed and who responded to the questionnaire mentioned using filters most often to identify randomised controlled trials and systematic reviews. The Information Specialists' Sub-Group (ISSG) Search Filters Resource was the most used source to find a filter, and over half of respondents relied on the filters available in databases they were searching. Searchers mentioned that having critical assessments of studies and user-friendly presentations of performance data available would help in choosing filters. Having filters available in a central location was also considered valuable.

Scientific summary

Background

The effective retrieval of relevant evidence is essential in the development of clinical guidance or health policy, the conduct of health research and the support of health-care decision-making. Whether the purpose of the evidence retrieval is to find a representative set of results to inform the development of an economic model or to find extensive evidence on the clinical effectiveness or cost-effectiveness of a health-care intervention, retrieval methods need to be appropriate, efficient within time and cost restraints, consistent and reliable.

One tool that can be useful for effective retrieval is the search filter. Search filters are a combination of search terms designed to retrieve records about a specific concept, which may be a study design, such as randomised controlled trials (RCTs), outcomes such as adverse events, a population such as women or a disease or condition such as cardiovascular disease. A methodological search filter is designed to capture the records of studies that have used a specific study design. Effective search filters may seek to maximise sensitivity (the proportion of relevant records retrieved), maximise precision (the proportion of retrieved records that are relevant) or optimise retrieval using a balance between maximising sensitivity and achieving adequate precision. Search filters can offer a standard approach to study retrieval and release searcher time to focus on developing other sections of the search strategy such as the disease concept.

Objectives

This project was funded to inform National Institute for Health and Care Excellence (NICE) methods development, but has wider application to efficient literature searching in support of evidence-based medicine in general. Its aim was to investigate the methods used to assess the performance of methodological search filters and explore what searchers require of search filters and what information searchers require to help them choose a search filter. We also explored systems and approaches for providing better access to relevant and useful performance data on methodological search filters, including developing suggested approaches to search filter performance measurement.

Our objectives were to identify and summarise:

- which performance measures for search filters are reported
- other performance measures reported in diagnostic test accuracy (DTA) studies and reviews
- different ways to present filter/test performance data to assist users in choosing which filters or tests to use
- evidence on how searchers choose search filters and what information they would like to receive to inform their choices
- evidence on how clinicians choose diagnostic tests.

The project website is at https://sites.google.com/a/york.ac.uk/search-filter-performance/ (accessed 22 August 2017).

Methods

We conducted a series of five literature reviews in 2010/11 into various aspects of search filter reporting and use and analogous activity in the field of DTA studies. The reviews informed the development of an

interview schedule, to learn how search filters are used by information professionals working for NICE and organisations affiliated to NICE, and also the development of a web-based questionnaire aimed at a wider audience of search experts in the area of search filters.

The literature reviews explored:

- what performance measures are reported for single studies of search filters and how are they presented (review A)
- what performance measures are reported when comparing a range of search filters and how the performance measures are synthesised (review B)
- what performance measures are reported in DTA studies and DTA reviews (review C)
- how searchers choose search filters (review D)
- how filter/test performance data are presented to assist users in choosing which filters or tests to use (reviews A, B and C)
- how clinicians or organisations choose diagnostic tests (review E).

Information professionals working for NICE, the NICE Collaborating Centres and NICE Evidence Review Groups were interviewed using a semistructured interview protocol.

A web-based questionnaire survey was developed to obtain information on searchers' knowledge of and use of search filters. The questions were based on findings from the reviews and the interviews. The questionnaire was advertised to seven e-mail discussion lists aimed at health librarians.

The reviews, interviews and questionnaire informed the development of suggested approaches to gathering and reporting search filter performance.

We acknowledge that there has been a regrettable delay between carrying out the project, including the searches, and the publication of this report, because of serious illness of the principal investigator.

Results

Review A

In total, 23 studies were identified in review A. In single studies reporting search filters:

- internal gold or reference standards were mostly derived by hand-searching journals
- filter validation was mostly carried out using internal validation
- sensitivity, precision and specificity were the most commonly used performance measures
- performance measures were most often presented in tables.

Review B

In total, 18 studies were identified in review B. In filter comparison studies:

- sensitivity, precision and specificity were the most commonly reported performance measures
- the highest sensitivity, highest precision and optimal/balanced filter strategies were most frequently reported
- methods reporting was limited in papers reporting the development of new search filters and comparison with existing filters
- the most frequently used method for reporting the results of filter performance comparisons was in tables, although graphs might be more useful.

Review C

In total, 47 studies were identified in review C. DTA studies and DTA reviews provided evidence that:

- studies should be carried out on a sample of patients who are representative of the target population and should use an appropriate reference standard
- sensitivity and specificity were the most commonly reported outcomes and are subject to spectrum bias
- predictive values are influenced by disease prevalence
- receiver operating characteristic curves present sensitivity and specificity pairs at different test thresholds
- the area under the curve gives an overall value of DTA
- health technology assessment organisations recommend that DTA studies should present 2 × 2 contingency tables, sensitivity and specificity pairs and likelihood ratio pairs
- several types of graphical presentation can be used to display DTA data but these had not been used extensively in the DTA literature
- poor-quality methods and reporting hinder the inferences that can be drawn from DTA studies.

Review D

No studies were identified that reported how searchers chose search filters.

Review E

Seven studies were identified that reported on factors that influenced clinicians' choice between diagnostic tests. They provided limited evidence suggesting that test performance is the main factor that informed choices. As a substantial proportion of clinicians have an inaccurate understanding of test performance parameters and how they should be applied, it might be the case that choices were being based on false assumptions.

Interviews

A total of 12 interviews were conducted, capturing the views of 16 information professionals.

The interviews revealed the wide range of searching tasks that are undertaken in the NICE context and the various points at which search filters can be used. The use of search filters seemed to be linked predominantly to reducing the numbers of retrieved records, introducing focus and assisting with searches that are focused on a single study type.

The Cochrane RCT and McMaster Hedges team filters were cited most often. Various methods were used to identify filters, with the most frequently mentioned resource being the Information Specialists' Sub-Group (ISSG) Search Filters Resource [Glanville J, Lefebvre C, Wright K. *ISSG Search Filter Resource*. York: The InterTASC Information Specialists' Sub-Group; 2008 (updated 2017). URL: https://sites.google.com/a/york.ac.uk/issg-search-filters-resource/home (accessed 22 August 2017).].

Interviewees' practices when using, adapting and reporting search filters were not uniform, possibly indicating an absence of accepted published formal guidance. Interviewees found it difficult to keep informed about search filter developments. When choosing filters, interviewees tried to make judgements around the relative sensitivity, specificity and precision of search filters but were conscious of factors such as time constraints and knowledge gaps that impeded this. Some interviewees requested more guidance on the best filters to use or chose filters based on the authorship of the filter. Some desire for standardisation or guidance within the NICE family was also expressed.

Questionnaire

In total, 90 individuals responded to the survey. About three-quarters of respondents said that they used search filters for extensive searches to inform guidelines or systematic reviews, with just over half saying that they would use them for rapid searches to answer brief questions and a similar number saying that they would use them for scoping searches to estimate the size of the literature on a topic.

The McMaster Hedges team was the most frequently reported source used to identify study design filters. Currently, respondents most frequently used search filters for RCTs and systematic reviews. The most frequently cited filters for a specific topic were the Cochrane RCT filters.

Just over half of the respondents reported that they generally use the in-built filters in database interfaces rather than typing in another filter. Once they had found a search filter, just over half of respondents reported that they sometimes amend the filter. Nearly all of those respondents who amended search filters tested the effect of the amendment by either comparing the results with and without the filter amendment or determining whether or not known relevant papers had been identified. Three-quarters of respondents documented their amendments when they wrote up the searches, using diverse approaches.

Information on search filter performance measures such as validation, sensitivity and precision, a description of the filter and the results of their own testing had helped respondents to choose between filters.

The main factors that would make choosing a filter easier were the availability of a critical appraisal or evaluation and more information on the effectiveness of the filter, what it does or what it provides, what it excludes, its limitations, when it was last updated, its advantages and disadvantages, its sensitivity and precision and what testing has been completed. Respondents wanted to be confident in the author/developer and the availability of the filter in a central location was important.

Conclusions

Studies of search filter development and comparison studies reached similar conclusions. Internal gold or reference standards were mostly derived by hand-searching journals. Internal rather than more rigorous external validation was more usually undertaken. The most commonly reported performance measures were sensitivity/recall, precision and specificity.

Filter performance comparison studies most commonly reported the highest sensitivity, highest precision and optimal/balanced filter strategies. These measures were generally presented in tables, with little use of other graphical options that might be more useful methods of presentation. Limited details about methods were reported and guidance in this area could be improved.

Guidance available on conducting and analysing the results of DTA studies is applicable to several aspects of search filter research. The identification of a representative sample of records, of sufficient size and using a standardised approach, will assist in producing robust and generalisable results. The greater use of graphical presentation might facilitate the dissemination and interpretation of results.

We did not identify any published research on how searchers choose search filters and were unable to draw conclusions. Furthermore, limited evidence was identified in the review of clinicians' decision-making, resulting in few insights into how clinicians or organisations choose diagnostic tests, which might have been transferable to the challenges of choosing search filters. Diagnostic test performance was the most frequent factor mentioned and is the main factor that is readily applicable to search filter choice. The other message that we identified is that providing additional explanatory information when reporting search filter performance might be necessary to ensure that searchers make choices based on an accurate understanding of test performance parameters.

The interviews and the questionnaire survey indicated that search filters are not appropriate for all searching tasks but are used mainly for reducing large results sets and assisting with searches that are focused on a single study type. Searchers use several key resources to identify search filters but may find choosing between filters challenging. Choosing filters might be aided by making information about filters less technical, offering ratings and providing more details about filter validation strategies and filter provenance.

The responses to the questionnaire provide many messages for search filter designers. Filter performance measures need to be signposted more clearly and succinctly to help searchers make better use of the available filters. Filter and website designers should present less information and ensure that performance information can be clearly identified. The provenance of filters is clearly important to some searchers but there are no established parameters to measure this confidence. Clear authorship labelling and the provision of detailed information to show the robustness of the development methods would not only assist users of filters but also help filter designers to achieve recognition for their filters. The convenience of having filters from well-established producers available within database interfaces encourages their use. A convenient filter may, however, not always be the best one for the task. Searchers need to know how to choose between a range of filters and need information on whether filters have been validated and how.

Recommendations for information retrieval practice

We recommend that:

- studies reporting search filter design and/or comparisons of search filter performance should explicitly report the methods and results to help searchers identify the most appropriate filter
- one or more gold or reference standards should be used for testing filter performance
- relative recall (RR) and hand-searching should be considered for the development of gold or reference standard(s) for filter development but caution should be exercised regarding the robustness of the original RR search
- search filters should be validated on gold or reference standards that are different from those from which they were developed (i.e. external validation)
- the size of the gold or reference standard(s) should be clearly stated and a sample size calculation presented to justify the size of the standard(s)
- when a filter has been translated for use in a different database and/or interface from that in which it was developed, this should be specifically reported
- results should be presented systematically, identifying clearly the best-performing filter for specific purposes (sensitive strategy, specific strategy, balanced strategy)
- tables of performance results should have a consistent format and order to enable information to be easily extracted
- additional reporting methods should be considered, including graphical options
- approaches such as those provided in this report should be considered regarding the use, adaptation and reporting of search filters.

Recommendations for research

Further research might include:

- the development and testing of filters for a wider range of study designs and other topics
- the development and testing of translations of filters for different databases and interfaces
- the development and testing of filters that are independent of indexing language
- a review of the performance measures reported and the methods of presentation used in methodological filter performance comparisons for study designs not included in this review
- studies to explore alternative methods of displaying performance results from comparisons of multiple methodological search filters
- explorations of methods for the numerical synthesis of the results of several filter performance comparisons.

Funding

The National Institute for Health Research (NIHR) Health Technology Assessment programme and Medical Research Council–NIHR Methodology Research Programme (grant number G0901496).

Chapter 1 Introduction

Background

The effective retrieval of relevant evidence is essential in the development of clinical guidance or health policy, the conduct of health research and the support of health-care decision-making. Whether the purpose of the evidence retrieval is to find a representative set of results to inform the development of an economic model or to find extensive evidence on the clinical effectiveness or cost-effectiveness of a health-care intervention, retrieval methods need to be appropriate, efficient within the time and cost restraints that exist, consistent and reliable.

One tool that can be useful for effective retrieval is the search filter. Search filters are a combination of search terms designed to retrieve records about a specific concept, which may be a study design, such as randomised controlled trials (RCTs), outcomes such as adverse events, a population such as women or a disease or condition such as cardiovascular disease. A methodological search filter is designed to capture the records of studies that have used a specific study design. Effective search filters may seek to maximise sensitivity (the proportion of relevant records retrieved), maximise precision (the proportion of retrieved records that are relevant) or optimise retrieval using a balance between maximising sensitivity and achieving adequate precision. Search filters can offer a standard approach to study retrieval and release searcher time to focus on developing other sections of the search strategy such as the disease concept.

Aims and objectives

This project was funded to inform National Institute for Health and Care Excellence (NICE) methods development by investigating the methods used to develop and assess the performance of methodological search filters, exploring what searchers require of search filters during the life of various types of projects and exploring what information searchers value to help them choose a search filter. We also explored systems and approaches for providing better access to relevant and useful performance data on methodological search filters, including developing suggested approaches to reliable and efficient search filter performance measurement.

Our objectives were to:

- identify and summarise the performance measures for search filters (single studies or performance reviews of a range of filters) that are reported
- identify and summarise other performance measures reported in diagnostic test accuracy (DTA) studies and DTA reviews
- identify and summarise ways to present filter/test performance data (e.g. graphs or tables) to assist users (searchers or clinicians) in choosing which filters or tests to use
- identify and summarise evidence on how searchers choose search filters
- identify and summarise evidence on how clinicians choose diagnostic tests
- understand better how searchers choose search filters and what information they would like to receive to inform their choices
- explore different ways to present search filter performance data for searchers and provide suggested approaches to presenting the performance data that searchers require
- develop suggested approaches for reliable and efficient measurement for search filter performance.

We acknowledge that there has been a regrettable delay between carrying out the project, including the searches, and the publication of this report, because of serious illness of the principal investigator. The searches were carried out in 2010/11.

Chapter 2 Methods

The research plan had several stages. It began with a series of five literature reviews into various aspects of search filter reporting and use. The reviews informed the development of an interview schedule and a web-based questionnaire (see *Appendix 1*). The reviews, interviews and questionnaire informed the development of suggested approaches to gathering and reporting search filter performance and a test website, on which we invite further feedback [see https://sites.google.com/a/york.ac.uk/search-filter-performance/ (accessed 22 August 2017)].

Reviews

The research was grounded in a series of five reviews. We conducted two reviews on how the performance of methodological search filters has been measured, in single studies and also in studies comparing the performance of search filters. In a third review we sought to find inspiration and synergies in the DTA literature by reviewing the literature on diagnostic test reporting and included an exploration of the potential relevance of performance measures used in DTA studies. Search filters are analogous to diagnostic tests, being designed to distinguish relevant records from irrelevant records, and the performance of search filters and diagnostic tests is reported using similar measures, such as sensitivity and specificity. A fourth review sought reports on how searchers make choices about filters based on the information presented to them and a fifth review sought to identify any information on how clinicians make choices about diagnostic tests to gain insights into how searchers do or might in the future be encouraged to make choices about search filters.

The reviews were informed by literature searches conducted in databases in a number of disciplines including information science. Further information about the searches can be found within each of the reviews described later in this chapter and the search strategies are all included in the relevant appendices. The sources searched were:

- The Cochrane Library
- EMBASE
- European network for Health Technology Assessment (EUnetHTA)
- health technology assessment (HTA) organisation websites
- Health Technology Assessment international (HTAi) Vortal
- Inter Technology Appraisal Support Collaboration (InterTASC) Information Specialists' Sub-Group (ISSG) Search Filters Resource
- Library and Information Science Abstracts (LISA)
- MEDLINE
- PsycINFO.

The reviews were conducted to reflect the project objectives, which were to determine:

- what performance measures are reported for single studies of search filters and how they are presented (review A)
- what performance measures are reported when comparing a range of search filters and how the performance measures are synthesised (review B)
- what performance measures are reported in DTA studies and DTA reviews (review C)
- how searchers choose search filters (review D)
- how filter/test performance data are presented (e.g. text, graphs, tables, graphics) to assist users (searchers or clinicians) in choosing which filters or tests to use (reviews A, B and C)
- how clinicians or organisations choose diagnostic tests (review E).

Interviews and questionnaire

The objective of the reviews was to identify information about:

- performance measures in use
- the presentation of performance measures
- how searchers and clinicians choose search filters or diagnostic tests.

The next stage, consisting of two phases (semistructured interviews and a questionnaire survey), was to ascertain which search filter performance measures were deemed to be the most important by searchers for informed decision-making. We sought to gain information on how search filter performance information could most usefully be presented to assist decisions and whether or not there is scope for performance information to be obtained as part of routine project work.

Phase 1: semistructured interviews

As this project was funded to inform NICE methods development, the involvement of NICE staff was central to it. We contacted NICE information specialists and project managers and offered them the opportunity to participate in the project. Each interview, which was recorded, lasted for no more than 45 minutes. Once the interview time and date were agreed, confirmation details (date, time, length of interview and interviewer details), along with a topic guide and assurance of anonymity, were sent to each interviewee. After each interview, an e-mail containing a summary of the key points raised during the interview was sent to each interviewee, who was offered the opportunity to check the notes for accuracy and add any additional points that may have occurred to him or her after the interview had ended.

Phase 2: questionnaire survey

Information from the literature reviews and the interviews was used to inform the design and content of a web-based questionnaire. NICE information specialists and project managers were invited to complete the questionnaire but it was also used to collect the views of the wider (national and international) systematic review, HTA and guidelines information community. This information community is well networked and was reached via e-mail lists, as described in *Chapter 4* (see *Questionnaire methods*).

Presentation of filter information

Information from the reviews and interview and questionnaire responses was used to develop suggested approaches to measuring search filter performance.

We also developed a series of pilot formats for presenting search filter performance information. With the approval of the authors, some of the data from the Cochrane methodology review of the performance of search filters in identifying DTA studies,[1,2] which at the time of the project was not yet published, was used to populate the pilot formats.

Performance tests, reports and performance resource

We developed a prototype web resource (using content management systems available at the University of York) to present performance data and to facilitate feedback and comments from NICE staff and others from within the evidence synthesis information community. Without prejudging users' requirements or the results of the research, the performance resource presented a matrix of information showing how well published search filters perform for specific study designs in different clinical specialties and with different user preferences for measures such as sensitivity or precision.

Based on the suggested approaches, we developed performance tests and performance reports, which were uploaded onto the project website. We also developed detailed procedures with the intention of assisting researchers to conduct and report future performance tests. We considered that if we could ascertain that users valued information in a specific format then we could try to develop suggested approaches to promoting these methods. The intention was to develop user-friendly tools for the future and to explore options to make these tools widely available.

Performance measures for methodological search filters (review A)

Introduction

Although there are a large number of search filters in existence, many have been developed pragmatically and have not undergone validation. Even for those search filters that have been validated, few have been validated beyond the data in the original publication. This method is described as internal validation and is a less rigorous approach than external validation, in which a filter is tested using a different gold standard from the one used to develop the filter. External validation provides an independent assessment of filter performance and gives a better indication of how a filter is likely to perform in the real world.

Selection of a search filter will depend on the particular searching task and on the performance of the search filter. Thus, it is important to report performance measures for search filters. There are a few tools available that can be used to assess or appraise search filters and these can help in the selection of search filters for specific tasks.[3–5]

The aim of this review was to look at the performance measures that are reported for search filters (single studies) and how they are presented. Single studies were defined as those in which a new search filter (or series of filters) was developed, or a search filter was revised, and in which performance measures of the search filter(s) were also reported.

The objectives of the review were to:

- identify and summarise the methods used to develop and validate search filters
- identify and summarise the performance measures used in single studies of search filters
- describe how these performance measures are presented.

Methods

Identification of studies

Studies were identified from the ISSG Search Filters Resource.[6] The ISSG Search Filters Resource is a collaborative venture to identify, assess and test search filters designed to retrieve health-care research by study design. It includes published filters and ongoing research on filter design, research evaluating the performance of filters and articles providing a general overview of search filters. At the time of this project, regular searches were being carried out in a number of databases and websites, and tables of contents of key journals and conference proceedings were being scanned to populate the site. Researchers working on search filter design are encouraged to submit details of their work. The 2010 update search carried out by the UK Cochrane Centre to support the ISSG Search Filters Resource website was also scanned to identify any relevant studies that were not included on the website at that time.

We acknowledge that there has been a regrettable delay between carrying out the project, including the searches, and the publication of this report, because of serious illness of the principal investigator. The searches were carried out in 2010/11.

Inclusion criteria

The review included studies that reported the development and evaluated the performance of methodological search filters for health-care bibliographic databases. For pragmatic reasons, the review specifically focused on studies that developed and evaluated methodological search filters for economic evaluations, DTA studies, systematic reviews and RCTs. These study types are the ones most commonly used by organisations such as NICE to underpin their decision-making when producing technology appraisals and economic evaluations of health-care technologies and subsequent clinical guidelines. Publications prior to 2001 were excluded partly for pragmatic reasons but also because during this period search filters tended to be derived by subjective methods and because some of the filters had subsequently been updated or were now out-of-date because of changes in database indexing.

Exclusion criteria

Studies were excluded from the review if they:

- were available only in abstract form (e.g. conference abstracts)
- did not develop or revise a search filter
- did not report details of the methods used in developing the search filter
- did not evaluate search filter performance
- were published before 2001.

Data extraction

Data were extracted from selected studies using a standardised data extraction form to identify information regarding gold/reference standards, filter development/validation and performance measures reported.

Results

Fifty-eight studies were identified from the ISSG Search Filters Resource. After applying the outlined inclusion and exclusion criteria, 23 studies were identified for inclusion in the review.[7–29] Details from the included studies, grouped according to type of methodological search filter (economic, diagnostic, systematic review and RCT), are provided in *Tables 1–4*.

Of the 35 studies excluded, 19 were rejected because they were published before 2001. The reasons why the remaining 16 studies were excluded are presented in *Table 5*.

Study details

Three studies included analyses of more than one search filter type: one study[12] included details of a diagnostic filter and a secondary (systematic review) filter and two studies[16,21] included details of both systematic review and RCT search filters. Thus, there were two studies examining economic search filters, seven studies examining diagnostic search filters, seven studies examining systematic review search filters and 10 studies examining RCT search filters.

The majority of the studies ($n = 14$)[8–10,12–14,17–19,22,23,26,27,29] addressed the development of search filters for use with MEDLINE, 10 for the Ovid platform,[8,9,13,14,17,19,22,23,27,29] three for PubMed[12,18,26] and one for DataStar.[10] Six studies developed search filters for the EMBASE database,[7,11,15,20,24,28] four for the Ovid platform,[7,15,20,28] one for DataStar[11] and one that used three different platforms (DataStar, Dialog and Ovid).[24] The remaining three studies developed search filters for the Cumulative Index to Nursing and Allied Health Literature (CINAHL),[21] PsycINFO[16] and the Latin American and Caribbean Health Sciences Literature (LILACS) database[25] respectively. The CINAHL and PsycINFO search filters used the Ovid platform whereas the LILACS database was searched using an internet interface.

TABLE 1 Review A: included studies – economic search filter studies

Reference	Database/platform	Gold standard to derive/report filter performance (internal)	Filter development	Filters tested	Performance measures reported (presentation)	Gold standard to report external validation	External validation measures
[a]McKinlay 2006[7]	EMBASE (Ovid)	Hand-search of 55 journals for publication year 2000 ($n = 183$ for costs; $n = 31$ for economics). Articles were assessed by six research assistants; inter-rater agreement previously established as > 80%	Index terms and text words from clinical studies and advice sought from clinicians and librarians. Terms with individual sensitivity of > 25% and specificity of > 75% were incorporated into the development of the filters. Terms were combined with Boolean OR	Six single terms and six combinations of terms were reported (three each for costs and economics): 1. Best specificity (with sensitivity of ≥ 50%) 2. Best sensitivity (with specificity of ≥ 50%) 3. Best optimised (based on the smallest absolute difference between sensitivity and specificity)	Sensitivity, specificity, precision, accuracy, confidence intervals reported (tables)	None	No
[a]Wilczynski 2004[8]	MEDLINE (Ovid)	Hand-search of 68 journals for publication year 2000 ($n = 199$ for costs, $n = 23$ for economics). Articles were independently assessed by two research assistants and disagreements were resolved by a third independent assessment	Subjective – index terms and text words from clinical studies and advice sought from clinicians and librarians. Terms with individual sensitivity of > 25% and specificity of > 75% were incorporated into development of the filters. Terms were combined with Boolean OR	Nine combinations of terms were reported (five for costs and four for economics): 1. Best sensitivity (with specificity of ≥ 50%) 2. Best specificity (with sensitivity of ≥ 50%) 3. Best optimised (based on the smallest absolute difference between sensitivity and specificity)	Sensitivity, specificity, precision (tables)	None	No

a Studies by the McMaster Hedges team.

TABLE 2 Review A: included studies – diagnostic search filter studies

Reference	Database/platform	Gold standard to derive/report filter performance (internal)	Filter development	Filters tested	Performance measures reported (presentation)	Gold standard to report external validation	External validation measures
Astin 2008[9]	MEDLINE (Ovid)	Derivation set: hand-search of six journals for publication years 1985, 1995 and 1988 ($n = 333$). Articles were assessed independently by three researchers and discrepancies were resolved by discussion	Candidate terms from previously published strategies and MeSH and text words from derivation set MEDLINE records. Terms were added sequentially beginning with terms with the highest PPV and at each step adding the term that retrieved the largest proportion of additional derivation set records. The steps were repeated until the highest sensitivity was achieved	One filter tested. Separate filter for retrieving imaging studies developed	Sensitivity, specificity, PPV, confidence intervals reported (tables)	Validation set: hand-search of six journals for the publication year 2000 ($n = 186$)	Sensitivity, specificity, PPV, confidence intervals reported (tables)
Bachmann 2002[10]	MEDLINE (DataStar)	Hand-search of four journals for publication year 1989 ($n = 83$). Articles were assessed independently by two researchers	Word frequency analysis of all words in MEDLINE records, excluding those not semantically associated with diagnosis. The 20 terms with the highest individual sensitivity × precision score plus MeSH exp "sensitivity and specificity" were combined with OR in a stepwise fashion into a series of strategies and were performance tested	Two filters tested	Sensitivity, precision, NNR, confidence intervals reported (tables)	Hand-search of same four journals for publication year 1994 ($n = 53$) and four different journals for publication year 1999 ($n = 61$)	Sensitivity, precision, NNR, confidence intervals reported (for 1994 data) (tables)

Reference	Database/platform	Gold standard to derive/report filter performance (internal)	Filter development	Filters tested	Performance measures reported (presentation)	Gold standard to report external validation	External validation measures
Bachmann 2003[11]	EMBASE (DataStar)	Hand-search of four journals for publication year 1999 by one researcher, 10% independently assessed by second researcher ($n = 61$)	Word frequency analysis of all words in EMBASE records, excluding those not semantically associated with diagnosis. The 10 terms with the highest individual sensitivity × precision score were combined with OR into a series of strategies and were performance tested	Eight filters tested, three filters recommended: 1. Highest sensitivity 2. High sensitivity + 'reasonable' precision 3. High precision + 'reasonable' sensitivity	Sensitivity, precision, NNR, confidence intervals reported (tables)	None	No
Berg 2005[12]	MEDLINE (PubMed)	PubMed search carried out on 25 November 2002 of cancer-related fatigue using NLINKS-EBN matrix search strategies ($n = 238$). Articles were assessed by two reviewers. Inter-rater reliability 0.71	Terms from the PubMed clinical queries diagnosis filter. Additional terms from MeSH and text terms from gold standard records and additional search filters. Terms were tested to see if they fulfilled one inclusion criterion including having individual sensitivity of > 5% and specificity of > 95%. Terms were combined with OR until sensitivity was maximised	Two filters tested: 1. Highest sensitivity 2. Highest specificity Separate filters to identify secondary data were also developed	Sensitivity, specificity, NNR, LR+ values (tables)	None	No
[a]Haynes 2004[13]	MEDLINE (Ovid)	Hand-search of 161 journals for publication year 2000 ($n = 147$). Articles were assessed by six research assistants. Inter-rater agreement was previously established as > 80%	Index terms and text words from clinical studies and advice sought from clinicians and librarians. Terms with individual sensitivity of > 25% and specificity of > 75% were incorporated into development of the filters. Tested combining terms with OR	Three single terms and nine combinations of terms reported: 1. Best sensitivity (with specificity of > 50%) 2. Best specificity (with sensitivity of > 50%) 3. Best optimised (based on smallest absolute difference between sensitivity and specificity)	Sensitivity, specificity, precision, accuracy, confidence intervals reported (tables)	None	No

continued

TABLE 2 Review A: included studies – diagnostic search filter studies (*continued*)

Reference	Database/platform	Gold standard to derive/report filter performance (internal)	Filter development	Filters tested	Performance measures reported (presentation)	Gold standard to report external validation	External validation measures
Vincent 2003[14]	MEDLINE (Ovid)	Reference set: studies included in 16 systematic reviews of diagnostic tests for deep-vein thrombosis and indexed in MEDLINE ($n = 126$ published from 1969 to 2000). Authors note that the reference set excluded many high-quality articles	(a) Identified terms from five existing strategies and added two text terms and MeSH terms commonly used in DTA (b) Excluded general MeSH terms from (a) (c) Reference set records not retrieved by (b) were examined to identify additional text and MeSH terms	Three filters tested. One filter was recommended as 'more balanced' with high sensitivity and improved precision	Sensitivity (table) (data available to calculate precision)	None	No
[a]Wilczynski 2005[15]	EMBASE (Ovid)	Hand-search of 55 journals for publication year 2000 for methodologically sound diagnostic studies ($n = 97$). Articles were assessed by six research assistants. Inter-rater agreement was previously established as > 80%	Index terms and text words from clinical studies and advice sought from clinicians and librarians. Terms with an individual sensitivity of > 25% and specificity of > 75% were incorporated into the development of the filters. Tested out combining terms with OR	In total, 6574 strategies were tested. Three single terms and five combinations of terms were reported: 1. Best sensitivity (with specificity of ≥ 50%) 2. Best specificity (with sensitivity of ≥ 50%) 3. Best optimised (based on the smallest absolute difference between sensitivity and specificity)	Sensitivity, specificity, precision, accuracy, confidence intervals reported (tables)	None	No

LR+, positive likelihood ratio; MeSH, medical subject heading; NLINKS-EBN, Language in Nursing Knowledge Systems – Evidence Based Nursing; NNR, number needed to read; PPV, positive predictive value.
a Studies by the McMaster Hedges team.

TABLE 3 Review A: included studies – systematic review search filter studies

Reference	Database/platform	Gold standard to derive/report filter performance (internal)	Filter development	Filters tested	Performance measures reported (presentation)	Gold standard to report external validation	External validation measures
Berg 2005[12]	MEDLINE (PubMed)	PubMed search carried out 25 November 2002 on cancer-related fatigue using NLINKS-EBN matrix search strategies ($n = 238$). Articles were assessed by two reviewers. Inter-rater reliability 0.55	Terms from the PubMed clinical queries systematic review filter. Additional terms from MeSH and text terms from gold standard records and additional search filters. Terms were tested to see if they fulfilled one inclusion criterion including having an individual sensitivity of > 5% and specificity of > 95%. Terms were combined with OR until sensitivity was maximised	Numerous filters tested – results reported only for the best filter, which had high sensitivity and high specificity. Separate filters to identify diagnostic tests were also developed	Sensitivity, specificity, NNR, LR+ values (tables)	None	No
[a]Eady 2008[16]	PsycINFO (Ovid)	Hand-search of 64 journals for publication year 2000 ($n = 58$). Articles were assessed by six research assistants. Inter-rater agreement was previously established as > 80%	Index terms and text words from clinical studies and advice sought from clinicians and librarians. Terms with individual sensitivity of > 10% and specificity of > 10% were incorporated into the development of the filters. Tested out combining terms with OR	One single term and four combinations of terms reported: 1. Best sensitivity (keeping specificity at ≥ 50%) 2. Best specificity (keeping sensitivity at ≥ 50%) 3. Best optimisation of sensitivity and specificity (based on the lowest possible difference between sensitivity and specificity)	Sensitivity, specificity, precision, accuracy, confidence intervals reported (tables)	None	No

continued

TABLE 3 Review A: included studies – systematic review search filter studies (*continued*)

Reference	Database/platform	Gold standard to derive/report filter performance (internal)	Filter development	Filters tested	Performance measures reported (presentation)	Gold standard to report external validation	External validation measures
[a]Montori 2005[17]	MEDLINE (Ovid)	Derivation set: hand-search of 10 journals for publication year 2000 ($n = 133$ used to test strategies). Internal validation set: validation data set excluding CDSR ($n = 332$ used to validate strategies). Articles were assessed by six research assistants. Inter-rater agreement was previously established as > 80%	Index terms and text words from clinical studies and advice sought from clinicians and librarians. Terms with individual sensitivity of > 50% and specificity of > 75% were incorporated into the development of the filters. Tested out combining terms with OR	Five single terms reported: best sensitivity (with specificity ≥ 50%), best specificity (with sensitivity of ≥ 50%), best precision (based on sensitivity of ≥ 25% and specificity of ≥ 50%). Two combination strategies maximising sensitivity and minimising the difference between sensitivity and specificity. Four combination strategies maximising precision	Sensitivity, specificity, precision, confidence intervals reported (tables)	Validation dataset: hand-search of 161 journals for publication year 2000 ($n = 753$)	Sensitivity, specificity, precision, confidence intervals reported (tables)
Shojania 2001[18]	MEDLINE (PubMed)	None	Relevant publication types ('meta-analysis', 'review', 'guideline') plus title and text words typically found in systematic reviews	One filter tested against two external gold standards and also applied to three clinical topics (screening for colorectal cancer, thrombolytic therapy for venous thromboembolism and treatment of dementia)	No	Sensitivity: 1. Sample of 100 records from DARE 2. 103 reviews identified from hand-searching the American College of Physicians Journal Club covering 1999 to September/October 2000 PPV: 3. MeSH search for three clinical topics and results screened for systematic reviews	Sensitivity, confidence intervals reported (tables); PPV, confidence intervals reported (tables)

Reference	Database/platform	Gold standard to derive/report filter performance (internal)	Filter development	Filters tested	Performance measures reported (presentation)	Gold standard to report external validation	External validation measures
White 2001[19]	MEDLINE (Ovid)	Hand-search of five journals for publication years 1995 and 1997 (quasi-gold standard, $n = 110$). Articles were assessed independently by two experienced researchers. Two sets for comparison: 1. 110 reviews that did not meet the criteria for a systematic review 2. 125 non-review reports The three data sets were matched for subject and split into a test set ($n = 256$, 75%) and a validation set ($n = 89$, 25%)	Textual analysis of quasi-gold standard test set records. MeSH and publication type analysed for each of three test sets. A total of 38 terms were analysed by discriminant analysis to determine which best distinguished between the three sets of records	Five models (filters) were tested on the full test set	Sensitivity, specificity, precision (tables)	One model was tested on the validation set. All models were tested in a 'real-world' scenario using Ovid MEDLINE on CD-ROM from 1995 to 1998 (and compared with two previously published strategies)	Sensitivity, precision (discussed in text), sensitivity, precision (table)
[a]Wilczynski 2007[20]	EMBASE (Ovid)	Hand-search of 55 journals for publication year 2000 ($n = 220$). Articles were assessed by six research assistants. Inter-rater agreement was previously established as > 80%	Index terms and text words from clinical studies and advice sought from clinicians and librarians. Terms with individual sensitivity of > 25% and specificity of > 75% were incorporated into the development of filters. Tested out combining terms with OR	Two single terms and four combinations of terms reported: 1. Best sensitivity (with specificity of ≥ 50%) 2. Best specificity (with sensitivity of ≥ 50%), best optimised (based on smallest absolute difference between sensitivity and specificity)	Sensitivity, specificity, precision, accuracy, confidence intervals reported, (tables)	None	No

continued

TABLE 3 Review A: included studies – systematic review search filter studies (*continued*)

Reference	Database/platform	Gold standard to derive/report filter performance (internal)	Filter development	Filters tested	Performance measures reported (presentation)	Gold standard to report external validation	External validation measures
[a]Wong 2006[21]	CINAHL (Ovid)	Hand-search of 75 journals for publication year 2000 ($n = 127$). Articles were assessed by six research assistants. Inter-rater agreement was previously established as > 80%	Index terms and text words from clinical studies and advice sought from clinicians and librarians. Terms with individual sensitivity of at least 10% and specificity of at least 10% were incorporated into development of the filters. Tested out combining terms with OR	Three single terms and four combinations of terms were reported: 1. Best sensitivity (with specificity of ≥ 50%) 2. Best specificity (with sensitivity of ≥ 50%) 3. Best optimised (based on the smallest absolute difference between sensitivity and specificity)	Sensitivity, specificity, precision, accuracy confidence intervals reported, (tables)	None	No

CDSR, Cochrane Database of Systematic Reviews; DARE, Database of Abstracts of Reviews of Effects; LR+, positive likelihood ratio; MeSH, medical subject heading; NLINKS-EBN, Language in Nursing Knowledge Systems – Evidence Based Nursing; NNR, number needed to read; PPV, positive predictive value.
a Studies by the McMaster Hedges team.

TABLE 4 Review A: included studies – RCT search filter studies

Reference	Database/platform	Gold standard to derive/report filter performance (internal)	Filter development	Filters tested	Performance measures reported (presentation)	Gold standard to report external validation	External validation measures
[a]Eady 2008[16]	PsycINFO (Ovid)	Hand-search of 64 journals for publication year 2000 ($n = 233$). Articles were assessed by six research assistants. Inter-rater agreement was previously established as > 80%	Index terms and text words from clinical studies and advice sought from clinicians and librarians. Terms with individual sensitivity of ≥ 10% and specificity of ≥ 10% were incorporated into development of the filters. Tested out combining terms with OR and used stepwise logistic regression	One single term and five combinations of terms were reported 1. Best sensitivity (keeping specificity at ≥ 50%) 2. Best specificity (keeping sensitivity at ≥ 50%) 3. Best optimisation of sensitivity and specificity (based on the lowest possible difference between sensitivity and specificity)	Sensitivity, specificity, precision, accuracy, confidence intervals reported (tables)	None	No
Glanville 2006[22]	MEDLINE (Ovid)	Database searches of MEDLINE and CENTRAL. Gold standard: randomly selected RCT records (1970, 1980, 1990, 2000) ($n = 1347$). Comparison group of randomly selected records of non-trials for the same years ($n = 2400$)	Frequency analysis of gold standard records to identify terms. Logistic regression analysis used to identify best-discriminating sets of terms in 50% of gold standard and comparison group records. Terms tested on remaining 50% of gold standard/comparison group records. Six search strategies were derived: two single-term and four multiterm strategies	Search strategies derived from 50% of the gold standard/comparison group records were tested on the remaining 50% of the records. No details given of performance measures	None	External gold standard: (a) MEDLINE records with MeSH "exp breast neoplasms" assessed as being RCTs ($n = 54$) (b) MEDLINE records from 2003 for four conditions identified as being RCTs ($n = 424$) External validation using six best-performing strategies	External gold standard: (a) Yield in identifying unindexed trials (discussed in text) (b) Sensitivity, precision (tables). One strategy with the highest sensitivity recommended

continued

TABLE 4 Review A: included studies – RCT search filter studies (*continued*)

Reference	Database/platform	Gold standard to derive/report filter performance (internal)	Filter development	Filters tested	Performance measures reported (presentation)	Gold standard to report external validation	External validation measures
[a]Haynes 2005[23]	MEDLINE (Ovid)	Hand-search of 161 journals for publication year 2000 ($n = 1587$); internal development set (60%) ($n = 930$); validation set (40%) ($n = 657$). Articles were assessed by six research assistants. Inter-rater agreement was previously established as > 80%	Index terms and text words from clinical studies and advice sought from clinicians and librarians. Terms with individual sensitivity of > 25% and specificity of > 75% were incorporated into the development of the filters. Tested out combining terms with OR and used stepwise logistic regression	Three single terms for high sensitivity, high specificity or optimised balance between sensitivity and specificity. Three combination strategies for highest sensitivity (specificity > 50%), three combination strategies for highest specificity (sensitivity > 50%), three combination strategies for highest accuracy (sensitivity > 50%), three combination strategies for optimising sensitivity and specificity (based on an absolute difference of < 1%). Best strategy for optimising trade-off between sensitivity and specificity when adding Boolean AND NOT. Best three combination strategies derived using logistic regression techniques	Sensitivity, specificity, precision, accuracy, confidence intervals reported (tables)	None	No

Reference	Database/platform	Gold standard to derive/report filter performance (internal)	Filter development	Filters tested	Performance measures reported (presentation)	Gold standard to report external validation	External validation measures
Lefebvre 2008[24]	EMBASE	Hand-search of two journals for publication years 1990 and 1994 ($n = 384$) were used to assess the performance of individual terms and select terms for further analysis. EMBASE records 1974–2005 (excluding those with corresponding MEDLINE record indexed as a RCT) and assessed as trials or not trials. This data set was used to combine and reject terms	MeSH terms from the MEDLINE Highly Sensitive Search Strategy were converted to Emtree where possible; additional Emtree terms and free text terms were also identified. Experts were consulted for further suggestions. Terms were tested against the internal gold standard records and those with an individual precision of > 40% and sensitivity of > 1% were selected and added sequentially to develop the filter. Terms with low cumulative precision were rejected	One filter	Cumulative sensitivity for each term, cumulative precision for each term and total (table)		

continued

TABLE 4 Review A: included studies – RCT search filter studies (*continued*)

Reference	Database/platform	Gold standard to derive/report filter performance (internal)	Filter development	Filters tested	Performance measures reported (presentation)	Gold standard to report external validation	External validation measures
Manríquez 2008[25]	LILACS (internet)	Hand-search of 44 Chilean journals for the publication years 1981–2004 indexed in LILACS ($n = 267$)	A total of 120 terms were identified from internal gold standard records. Terms with individual sensitivity of > 20% and specificity and accuracy of > 60% were included in two-term strategies. Terms in two-term strategies with sensitivity, specificity and accuracy of > 60% were combined to give three- or four-term strategies. All terms in three- to four-term strategies were combined to give a maximum sensitivity strategy. The final strategy excluded terms with 0% sensitivity and high specificity	The sensitivity, specificity and accuracy are given for 16 single terms, 23 two-term strategies and 13 three- or four-term strategies. Sensitivity and specificity are given for a 10-term strategy (A) and a final strategy (B) (B is derived from strategy A by excluding terms with a sensitivity of 0% and high specificity)	Sensitivity, specificity (figure). The figure contains the full search strategy and values for sensitivity and specificity	None	No
Robinson 2002[26]	MEDLINE (PubMed)	None	Adapted from a previous search filter (Cochrane Highly Sensitive Search Strategy). Three revisions to the original Cochrane RCT strategy. Strategies also translated for PubMed	Comparison of results retrieved by the original and revised strategies for both MEDLINE Ovid and PubMed	Number of additional relevant and non-relevant records retrieved by revisions	Cochrane CENTRAL records from 11 journals for 1998 ($n = 308$)	Sensitivity (discussed in text of article)

Reference	Database/platform	Gold standard to derive/report filter performance (internal)	Filter development	Filters tested	Performance measures reported (presentation)	Gold standard to report external validation	External validation measures
[b]Taljaard 2010[27]	MEDLINE (Ovid)	Hand-search of 78 journals for one randomly assigned year from 2000 to 2007 ($n = 162$). Subset initially examined independently by two reviewers – inter-rater reliability of 0.81	Frequency analysis of text from internal gold standard records was used to create a search strategy for identifying CRTs	Three filters were tested: simple – RCT.pt; sensitive – identified CRT terms combined using OR and then combined with RCT.pt using OR; precise: identified CRT terms combined using OR and then combined with RCT.pt using AND	Sensitivity, precision, 1 – specificity (fallout) (tables), NNR (discussed in text of article)	Seven systematic reviews of CRTs covering 1979–2005 ($n = 363$)	Sensitivity (table) (referred to as RR in the text)
[a]Wong 2006[21]	CINAHL (Ovid)	Hand-search of 75 journals for publication year 2000 ($n = 506$). Articles were assessed by six research assistants. Inter-rater agreement was previously established as > 80%	Index terms and text words from clinical studies and advice sought from clinicians and librarians. Terms with an individual sensitivity of at least 10% and specificity of at least 10% were incorporated into the development of the filters. Tested out combining terms with OR and used stepwise logistic regression	Three single terms and five combinations of terms were reported: (1) best sensitivity (with specificity of ≥ 50%), (2) best specificity (with sensitivity of ≥ 50%), (3) best optimised (based on the smallest absolute difference between sensitivity and specificity)	Sensitivity, specificity, precision, accuracy, confidence intervals reported (tables)	None	No

continued

TABLE 4 Review A: included studies – RCT search filter studies (*continued*)

Reference	Database/platform	Gold standard to derive/report filter performance (internal)	Filter development	Filters tested	Performance measures reported (presentation)	Gold standard to report external validation	External validation measures
[a]Wong 2006[28]	EMBASE (Ovid)	Hand-search of 55 journals for publication year 2000 ($n = 1256$). Articles were assessed by six research assistants. Inter-rater agreement was previously established as > 80%	Index terms and text words from clinical studies and advice sought from clinicians and librarians. Terms with individual sensitivity of > 25% and specificity > 75% were incorporated into development of the filters. Tested out combining terms with OR	Three single terms and four combinations of terms were reported: (1) best sensitivity (with specificity of ≥ 50%), (2) best specificity (with sensitivity of ≥ 50%), (3) best optimised (based on the smallest absolute difference between sensitivity and specificity)	Sensitivity, specificity, precision, accuracy, confidence intervals reported (tables)	None	No
Zhang 2006[29]	MEDLINE (Ovid)	None	Used existing filters and revisions of existing filters	Evaluated six filters: the top two phases of the Cochrane Highly Sensitive Search Strategy (SS123, SS12) and four revisions of this strategy (SS-crossover, SS-crossover studies, SS-volunteer, SS-versus)	No	A total of 61 reviews identified from the CDSR in 2003 that had used the Highly Sensitive Search Strategy to identify RCTs and provided details of the subject search	Sensitivity, precision, article read ratio, interquartile ranges reported (tables)

CDSR, Cochrane Database of Systematic Reviews; CENTRAL, Cochrane Central Register of Controlled Trials; CRT, cluster randomised trial; MeSH, medical subject heading; NNR, number needed to read; RR, relative recall.
a Studies by the McMaster Hedges team.
b CRT.

TABLE 5 Review A: excluded studies

Study identifier	Reason for exclusion	Type of filter
Abhijnhan 2007[30]	Did not develop and test a filter. Focus is on a comparison of database content/coverage	RCT
Almerie 2007[31]	Did not develop and test a filter. Focus is on a comparison of database content/coverage	RCT
Chow 2004[32]	Did not develop or revise a filter	RCT
	Methods used to develop filter not reported	
Corrao 2006[33]	Filter not evaluated against either internal or external gold standards	RCT
	No internal or external validation standards	
Day 2005[34]	Did not develop or test a RCT search filter. The search strategies derived were based on the condition and intervention of interest	RCT
de Freitas 2005[35]	Did not develop and test a filter	RCT
Devillé 2002[36]	This was a guideline for conducting diagnostic systematic reviews	Diagnostic
	No filter development or evaluation	
Eisinga 2007[37]	Did not develop or revise a filter	RCT
Kele 2005[38]	Did not develop and test a filter. Focus is on a comparison of database content/coverage	RCT
Kumar 2005[39]	Did not develop and test a filter. Focus is on a comparison of database content/coverage	RCT
McDonald 2002[40]	Did not develop and test a filter	RCT
Royle 2003[41]	Did not develop or revise a search filter	Economic
	Did not evaluate a search filter	
	Focus is on sources used for searching for economic studies	
Royle 2005[42]	Did not develop and test a filter	RCT
Royle 2007[43]	Methods used to develop filter not reported	RCT
Sassi 2002[44]	Methods used to develop search filter not reported	Economic
	No gold standard – comparator is an 'extensive search'	
Wilczynski 2009[45]	Focus is on the quality of indexing of systematic reviews and meta-analyses in MEDLINE	Systematic review

Internal gold standards

A reference standard is a set of relevant records against which a search filter's performance can be measured. In some studies the reference standard is used both to derive and to test a search filter. In these cases the standard is described as an internal standard.

Almost all of the studies used an internal standard to derive and/or validate the search filters. Only three of the 23 studies did not include an internal standard.[18,26,29] These studies tested the search filters against external standards (see *External standards*). Seventeen[7–11,13,15–17,19–21,23–25,27,28] of the 20 studies that included an internal standard had derived this standard by hand-searching journals. The number of journals searched ranged from 2 to 161. In the other three studies[12,14,22] the internal standards were generated by a PubMed subject-specific search or from studies included in a number of systematic reviews, or from a database search [MEDLINE and the Cochrane Central Register of Controlled Trials (CENTRAL)]. One other study[24] used a search of EMBASE as well as hand-searching of journals to derive an internal standard. The size of the gold or reference standards varied from 58 to 1587 records. In three studies, the reference standard was initially split into two, with one set used to derive the filter and the second set used to internally validate the performance.[17,19,23]

Inter-rater reliability in selecting studies for inclusion in the reference standard was assessed for almost all of the studies produced by the McMaster Hedges team[7,8,13,15–17,20,21,23,28] and exceeded 80% in every case. In one McMaster Hedges team study,[8] articles were independently assessed by two reviewers with disagreement being resolved by a third independent reviewer. Two studies quoted inter-rater reliabilities of 71%[12] and 81%[27] after articles were assessed by two reviewers. Two further studies[10,19] reported that articles were assessed by two reviewers, whereas one study[11] reported that articles were assessed by one reviewer with 10% of articles assessed by a second reviewer and one study[9] reported that articles were assessed by three researchers with discrepancies resolved through discussion. None of these studies reported values for inter-rater reliability. The remaining four studies that derived internal standards[14,22,24,25] did not describe how the studies were selected.

Identifying candidate terms and combining them to create filters

In the 20 studies with internal standards, the internal standard records were used as a source for the identification of candidate search terms. Ten of these studies[7,8,13,15–17,20,21,23,28] were carried out by the McMaster Hedges team and used essentially the same methodology for deriving search filters. This method involved the identification of index terms and text words from an internal standard of records as well as consultation with clinicians, librarians and other experts to add any other relevant terms. The individual terms identified were analysed for sensitivity and specificity and then terms with specified values of sensitivity and specificity were combined to create multiple-term search filters using the Boolean OR operator. The specified values for term inclusion varied for sensitivity and specificity from > 10% to > 75%. In one of the 10 studies[23] stepwise logistic regression was also used to try to optimise search filter performance. The use of logistic regression, however, did not result in better-performing search filters than those developed simply using the Boolean OR operator and therefore this approach was not used in any of the subsequent studies.

Another study[25] also identified terms from an internal standard and then combined terms with particular values for sensitivity, specificity and accuracy to derive multiple-term strategies to produce a maximally sensitivity strategy. Single terms with an individual sensitivity of > 20% and specificity and accuracy of > 60% were combined to give two-term strategies. Terms in the two-term strategies with sensitivity, specificity and accuracy of > 60% were then combined to give three- or four-term strategies. All terms in the three- and four-term strategies were then combined to give a maximally sensitivity strategy consisting of 10 terms. This final strategy was refined further by using the Boolean AND NOT operator to exclude single terms with zero sensitivity and high specificity. This increased the specificity of the final strategy while maintaining high sensitivity.

Five studies[10,11,19,22,27] used bibliographic software to undertake a more formal frequency analysis of the terms in the internal standard. Two of these studies[10,11] carried out word frequency analysis for all of the records in the internal standard and then created search strategies by combining those terms that had the highest scores as determined by multiplying the sensitivity and precision scores. Two studies[19,22] used textual analysis of the internal standard records followed by discriminant analysis using logistic regression to determine the best terms to be included in the search strategy. The fifth study[27] also used frequency analysis to identify candidate terms for building a search strategy.

Previously published filters were used as a source of terms for four studies.[9,12,14,24] These strategies were then further developed by adding extra medical subject heading (MeSH) and text terms identified from the internal standard records. In one of these studies[24] the MeSH terms were first translated from a MEDLINE strategy into Emtree terms before adding additional Emtree terms and free-text terms identified from the internal standard records. This study also consulted experts for further suggestions. Individual terms were tested against the internal standard and those with a precision of > 40% and sensitivity of > 1% were added sequentially to develop the filter. Astin *et al*.[9] also used the sequential addition of search terms to develop the search filter.

Internal validation performance measures

The performance of the search filters was tested against the gold or reference standard in 19 studies[7–17,19–21,23–25,27,28] to test internal validity. Nine studies[7,13,15–17,20,21,23,28] carried out by the McMaster Hedges team reported the results for single-term and combined-term search strategies, whereas the remaining study[8] from this team reported only the performance of combination-term strategies. Studies reporting single-term strategies included between one and six single-term strategies whereas the number of combination strategies reported varied between four and 14. The performance of strategies was usually reported in terms of high sensitivity, high specificity or optimised balance between sensitivity and specificity. The other nine studies[9–12,14,19,24,25,27] tested between one and eight filters, with some single-term strategies but mostly combination strategies. The focus of these search filters was to produce highly sensitive, highly specific or highly precise outcomes.

The performance measures reported for internal validation are presented in *Table 6*. Sensitivity was reported by all 19 studies, precision was reported by 16 studies and specificity was reported by 14 studies. Accuracy was reported by seven studies and the number needed to read (NNR) by four studies. Positive likelihood ratio (LR+) values and fall-out were each only reported in a single study. All of the performance measures were presented in tables with the exception of one study,[25] for which the results were presented in a figure that contained the full search strategy and values for sensitivity and specificity.

External standards

Nine of the 23 studies used external standards to test or validate the search filters that had been developed or revised.[9,10,17–19,22,26,27,29] For these studies, a reference standard that was different from the one used to derive the search filter was used. These studies included studies of diagnostic test, systematic review and RCT filters. Four studies[9,10,17,18] used hand-searching of journals to generate the external standard. The number of journals searched ranged from 1 to 161, resulting in between 53 and 332 records in the external standards. Two of these four studies[17,18] increased the numbers in the external standard by adding records from a search of either the Cochrane Database of Systematic Reviews (CDSR) or the Database of Abstracts of Reviews of Effects (DARE).

Four[22,26,27,29] of the other five studies that used external standards were of RCT search filters and one[19] was of a systematic review search filter. Two of these studies[27,29] identified records for their standards by searching systemic reviews (one searched 61 reviews from the CDSR[29] and one[27] searched seven systematic reviews of cluster RCTs). Another study[26] searched for records in 11 journals in the CENTRAL database, generating 308 references. In the remaining RCT search filter study[22] MEDLINE was searched to identify records that were assessed as being trials. In the study that examined a systematic review search filter[19] models were tested using a validation data set and against a 'real-world' scenario using Ovid MEDLINE on compact disc, read-only memory (CD-ROM). The validation data set had been created from a hand-search of five journals.

TABLE 6 Review A: performance measures – internal standards

Performance measure	Number of studies reporting the performance measure	Reference numbers of articles reporting the studies	Percentage of studies reporting the performance measure
Sensitivity	19	7–17,19–21,23–25,27,28	100
Specificity	14	7–9,12,13,15–17,19–21,23,25,28	74
Precision (or PPV)	16	7–11,13,15–17,19–21,23,24,27,28	84
Accuracy	8	7,13,15,16,20,21,23,28	42
NNR	4	10–12,27	21
LR+	1	12	5
Fall-out	1	27	5

PPV, positive predictive value.

The results of this hand-search had been split into an internal test set ($n = 256$, 75%) and an external validation set ($n = 89$, 25%).

External validation performance measures

The performance of the search filters was tested against external standards in nine studies.[9,10,17–19,22,26,27,29] The performance measures reported for external validation are presented in *Table 7*. All nine studies reported sensitivity and seven of the nine studies reported precision. Two studies[9,17] reported specificity and two[10,29] reported the NNR (described as 'article read ratio' in one article). Two studies[26,27] reported a single performance measure, that is, sensitivity only, three studies[18,19,22] reported two performance measures and four studies[9,10,17,29] reported three performance measures. The performance measures were again presented almost exclusively in tables, with one exception,[26] in which the performance measures were simply discussed in the text of the article.

Discussion

Methods used to develop and validate search filters

A total of 23 studies were included in this review. In the majority of these studies an internal gold or reference standard was used to develop the search filter by identifying candidate terms and assessing performance. The way in which terms were chosen for inclusion, however, and how the combinations were determined varied. The internal gold standards were mainly derived from journal hand-searches although a few were derived by other methods (from a database search or studies identified from systematic reviews). Ten of the studies were produced by the McMaster Hedges team and these all used the same method of search filter development, for example through consultation with experts and use of their internal gold or reference standard. Five other studies made use of statistical methods for filter development. The use of statistical methods helps to make the process more objective rather than depending on human expertise. In a few cases, the search filter was not developed using a gold standard or reference standard but was adapted from a previous search filter. Only nine studies undertook external validation, that is, validation against a standard that was different from the one used to develop the filter. As this provides an independent assessment of filter performance, it provides a more rigorous assessment and gives a better indication of how a filter is likely to perform in the real world.

Reported performance measures

Across the 23 studies included in the review, eight different performance measures were reported; however, as precision and positive predictive value (PPV) are equivalent, there were actually seven different performance measures. The performance measures used for internal and external validation and their frequency of use are listed in *Tables 6* and *7* respectively. The most frequently reported performance measures were sensitivity, precision and specificity respectively.

All studies reported sensitivity, reflecting the importance of this measure when determining the usefulness of a search filter. As the filters are used to identify relevant articles, it is important to measure the number of relevant articles retrieved by the filter compared with the total possible number of relevant articles.

TABLE 7 Review A: performance measures – external standards

Performance measure	Number of studies reporting the performance measure	Reference numbers of articles reporting the studies	Percentage of studies reporting the performance measure
Sensitivity	9	9,10,17–19,22,26,27,29	100
Specificity	2	9,17	22
Precision (or PPV)	7	9,10,17–19,22,29	78
NNR (article read ratio)	2	10,29	22

PPV, positive predictive value.

When carrying out a systematic review, in which it is important to identify as many relevant studies as possible, it makes sense to use a search filter with a high sensitivity value.

The performance measures of specificity and precision were the next most reported measures. It is important that a search filter rejects non-relevant articles and thus a high specificity is desirable. In a well-performing search filter a high specificity value would be desirable as well as a high sensitivity value, as there would not be much point in using a filter that retrieves lots of non-relevant articles as well as all of the relevant articles. The articles in the review often included search filters that were optimised for the best balance of sensitivity and specificity.

As precision measures the number of relevant articles as a proportion of all articles retrieved, the aim is to maximise the precision of a search filter. As sensitivity and precision are, however, inversely related, it is difficult to achieve both high sensitivity and high precision. The NNR is another way of reporting precision as it is calculated by dividing 1 by the precision value. This measure gives the number of articles that need to be read to find one relevant article and may, therefore, be more easily understood than precision, which is usually quoted as a percentage value.

The accuracy performance measure was used only in articles produced by the McMaster Hedges team. It provides a measure of the number of articles that are classified correctly as either relevant or non-relevant. The usefulness of this measure on its own, however, is unclear as a high accuracy value may be obtained when the specificity value is high but the sensitivity value is medium or low. In most cases the accuracy value is close to the specificity value and does not give an indication of the sensitivity value.

The other two performance measures that were found (LR+ and fall-out) each appeared in one article. These performance measures were reported in addition to sensitivity and either specificity or precision.

Presentation of performance measures

The most commonly used format for the presentation of performance measures used for single studies of search filters was tables. Only two studies of RCT filters did not present the performance measures in tables. One of these studies presented the search strategy and its performance measures in a figure whereas the other study simply discussed the performance measures in the text of the article. Thus, tables seem to be a popular and useful way of presenting performance measures. Often the results are ordered in tables according to one of the performance measures, for example sensitivity, thus making it easy to identify the most sensitive and the least sensitive search filter. The studies often presented the performance measures in a number of tables to allow ordering by different performance measures, for example tables ordered by sensitivity or specificity or precision. This makes it easier to select a search filter for a specific need, for example researchers involved in performing systematic reviews requiring very sensitive search filters could select the most sensitive search filters whereas busy clinicians who are simply looking for some relevant articles could select a filter with the highest precision.

Key findings

- Internal gold or reference standards were mostly derived by hand-searching of journals.
- Validation of filters was mostly carried out using internal validation.
- The most commonly used performance measures were sensitivity, precision and specificity.
- The majority of the studies presented performance measures in tables.

Measures for comparing the performance of methodological search filters (review B)

Reproduced with permission from Harbour *et al.*[46] © 2014 The authors. Health Information and Libraries Journal © 2014 Health Libraries Journal. *Health Information & Libraries Journal*, **31**, pp. 176–194.

Introduction

A variety of methodological search filters are already available to find RCTs, economic evaluations, systematic reviews and many other study designs. In principle, these filters can offer efficient, validated and consistent approaches to study identification within large bibliographic databases. Search filters, however, are an under-researched tool. Although there are many published search filters, few have been extensively validated beyond the data offered in the original publications.[47–49] This means that their performance in the real-world setting of day-to-day information retrieval across a range of search topics is unknown.[50] Furthermore, search filters are seldom assessed against common data sets, which makes a comparison of performance across filters problematic. Consequently, the use of search filters as a standard tool within technology assessment, guideline development and other evidence syntheses may be pragmatic rather than evidence based.[50,51]

As search filters proliferate, the key question becomes how to choose between them. The most useful information to assist search filter choice is likely to be performance data derived from well-conducted and well-reported performance tests or comparisons. Methods exist to test search filter performance and to build the performance picture, including reviews of search filter performance.[48,49,52–54] There is no formal guidance, however, on the best methods for testing filter performance, on which performance measures are valued by searchers and on which measures should ideally be reported to assist searchers in choosing between filters. The performance picture for filters across different disciplines, questions and databases is therefore largely unknown. Different performance measures are reported in studies describing search filters and the process whereby searchers choose a filter remains unclear.

The purpose of this review was to consider the measures and methods used in reporting the comparative performance of multiple methodological search filters.

Objectives

This review addressed the following questions:

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

Methods

Identification of studies

Studies were identified from the ISSG Search Filters Resource.[6] The ISSG Search Filters Resource is a collaborative venture to identify, assess and test search filters designed to retrieve health-care research by study design. It includes published filters and ongoing research on filter design, research evaluating the performance of filters and articles providing a general overview of search filters. At the time of this project, regular searches were being carried out in a number of databases and websites, and tables of contents of key journals and conference proceedings were being scanned to populate the site. Researchers working on search filter design are encouraged to submit details of their work. The 2010 update search carried out by the UK Cochrane Centre to support the ISSG Search Filters Resource website was also scanned to identify any relevant studies not at that time included on the website. We acknowledge that there has been a regrettable delay between carrying out the project, including the searches, and the publication of this report, due to serious illness of the principal investigator. The searches were carried out in 2010/2011.

Inclusion criteria

For the purpose of this review, methodological search filters were defined as any search filter or strategy used to identify database records of studies that use a particular clinical research method. A pragmatic decision was taken to include only studies comparing the performance of filters for RCTs, DTA studies, systematic reviews or economic evaluation studies. These study types are the ones most commonly used by organisations such as NICE to underpin their decision-making when producing technology appraisals and economic evaluations of health-care technologies and subsequent clinical guidelines.

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

Exclusion criteria

Studies were excluded from the review if they:

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

Data extraction and synthesis

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

Results

Twenty-one studies were identified as potentially meeting the inclusion criteria for this review based on titles and abstracts[2,10,14,15,17,19,22,23,25,33,48,49,55–63] Of these studies, 10 reported the development of one or more search filters, whose performance was then compared against the performance of existing filters[10,14,15,17,19,22,23,25,56,57] and 11 reported the comparative performance of existing filters.[2,33,48,49,55,58–63] On receipt of the full articles, three studies[55,60,62] were excluded from the review based on the criteria outlined in the methods section. The 18 included studies are listed in *Tables 8* and *9* and the excluded studies are listed in *Table 10*. No studies were identified that synthesised the results of several performance reports or reviewed the results of several syntheses.

Of the 18 studies included in the review:

- eight reported the performance of DTA search filters[2,10,14,15,48,49,57,58]
- five reported the performance of RCT filters[22,23,25,33,61]
- three reported the performance of systematic review filters[17,19,56]
- one reported the performance of filters for economic evaluations[59]
- one reported the performance of RCT and systematic review filters.[63]

The methodological filters evaluated in the included studies had been developed in a variety of interfaces including the interfaces to LILACS, Ovid, PubMed and SilverPlatter. Most studies, however, did not specify the interface used in the development of some or all of the filters being compared.[2,15,17,19,22,23,49,56–59,61]

TABLE 8 Review B: characteristics of the performance comparison studies included in this review[a]

Study	How were filters identified for comparison?	What study type was the filter designed to retrieve?	Total number of included filters (number of included filters developed by the author)	Database in which filters were tested
Bachmann 2002[10]	Published filters	DTA studies	2 (1)	MEDLINE
Boynton 1998[56]	Published filters	Systematic reviews	15 (11)	MEDLINE
Corrao 2006[33]	Published filters, author-modified strategy	RCTs	2	MEDLINE
Devillé 2000[57]	Published filters	DTA studies	5 (4)	MEDLINE
Doust 2005[58]	Published filters	DTA studies	5	MEDLINE
Glanville 2006[22]	Published filters	RCTs	12 (6)	MEDLINE
Glanville 2009[59]	Websites, contact with experts	Economic evaluations	22	MEDLINE and EMBASE
Haynes 2005[23]	Websites, published filters	RCTs	21 (2)	MEDLINE
Leeflang 2006[48]	Database search	DTA studies	12	MEDLINE
Manríquez 2008[25]	Published filters	RCTs	2 (1)	LILACS
McKibbon 2009[61]	Database search, websites, published filters	RCTs	38	MEDLINE
Montori 2005[17]	Published filters	Systematic reviews	10 (4)	MEDLINE
Ritchie 2007[49]	Database search, contact with experts, published filters	DTA studies	23	MEDLINE
Vincent 2003[14]	Database search, websites	DTA studies	8 (3)	MEDLINE
White 2001[19]	Published filters	Systematic reviews	7 (5)	MEDLINE
Whiting 2011 (online 2010)[2]	Contact with experts, database search, published filters	DTA studies	22	MEDLINE
Wilczynski 2005[15]	Published filters	DTA studies	4 (2)	EMBASE
Wong 2006[63]	Published filters	RCTs and systematic reviews	13	MEDLINE and EMBASE

a Full details provided in *Table 9*.

This absence of detail was particularly common in studies in which performance comparison was secondary to the development of one or more new filters.[15,17,19,22,23,56,57]

Fourteen studies compared the performance of filters in MEDLINE (various interfaces).[2,10,14,17,19,22,23,33,48,49,56–58,61] Two studies tested filters in MEDLINE and EMBASE.[59,63] One study only tested EMBASE filters[15] and one study compared filters in LILACS.[25] Seven of the eight studies comparing DTA filters used MEDLINE to test performance, although the interface used varied.[2,10,14,48,49,57,58]

Studies included in the review used a variety of methods to identify relevant filters for comparison, including database searches,[2,14,48,49,61] consulting relevant websites[14,23,59,61] and contacting experts in the field.[2,49,59] Ten studies used other methods of identifying filters such as using studies that they already knew about or studies that they had conducted themselves.[2,10,17,22,23,49,57,58,61,63] Five studies did not provide explicit details on how the filters for testing were identified.[15,19,25,33,56]

The number of filters compared in a single study ranged from 2 to 38. DTA study and RCT filters were the most common filters compared and systematic review and economic evaluation filters were the least common.

TABLE 9 Review B: table of included studies

Study	Filters included	Tested in	Identification of filters	Filter translation	Gold standard	Method of testing	Measures reported
Studies reporting on the comparative performance of published filters							
Corrao 2006[33]	Two RCT filters	PubMed	PubMed Clinical Queries specific therapy filter and authors' modified version: addition of term "randomised [Title/Abstract]"	Not required	None	Retrieved citations 'formally checked' to confirm RCT study design	Number retrieved that were confirmed RCTs, precision, retrieval gain (absolute and percentage)
Doust 2005[58]	Five DTA study filters	MEDLINE (WebSpirs)	Published strategies for diagnostic systematic reviews (no further details given)	Reports conversion from PubMed to MEDLINE (WebSpirs) for one filter. Reproduced terms used for all filters but did not discuss translation	Included studies from two systematic reviews. Studies identified from MEDLINE search using Clinical Queries diagnostic filter and reference check – 53 records	Filter terms, complete filter and filter plus original subject searches for reviews. Did not report date searched	Sensitivity/recall, precision
Glanville 2009[59]	14 MEDLINE economic evaluation study filters; eight EMBASE economic evaluation study filters	MEDLINE and EMBASE (Ovid)	Consulted websites and experts	Strategies adapted for Ovid 'as necessary' and reported in supplementary table	Records coded as economic evaluations in NHS EED (2000, 2003, 2006) and indexed in MEDLINE or EMBASE – MEDLINE 1955 records, EMBASE 1873 records	Filters run in MEDLINE and EMBASE for the same years as the gold standard with and without exclusions (animal studies and publication types unlikely to yield economic evaluations)	Sensitivity, precision

continued

TABLE 9 Review B: table of included studies (*continued*)

Study	Filters included	Tested in	Identification of filters	Filter translation	Gold standard	Method of testing	Measures reported
Leeflang 2006[48]	12 DTA study filters	PubMed	MEDLINE, EMBASE and Cochrane Methodology Register searches. When multiple filters were reported selected highest sensitivity, highest specificity and highest accuracy filters according to the original author(s)	Strategies adapted for PubMed. Translations reported in full	Included studies from 27 systematic reviews – 820 records	Filters run against PubMed records. Replicated original searches for six reviews with the addition of filters and using the same time frame	NNR, proportion of original articles missed, average proportion of retrieved and missed gold standard records per filter (bar chart), proportion of articles not identified per year (graph)
McKibbon 2009[61]	38 RCT filters	MEDLINE (Ovid)	Database (PubMed) searches, web searches, consulted websites, reviewed bibliographies, personal files	Strategies translated for Ovid. Translated filters reported in appendix	Hand-searching of 161 journals in 2000 – 1587 records of RCTs	Filters run in Clinical Queries Hedges database (49,028 MEDLINE records from hand-searched journals)	Sensitivity/recall, precision, specificity, confidence intervals reported
Ritchie 2007[49]	23 DTA study filters	MEDLINE (Ovid)	MEDLINE search, personal files, contacted experts	Reports one strategy translated from SilverPlatter to Ovid	Included studies from one review indexed in MEDLINE – 160 records	Replicated original review search (noted small discrepancy in results) with addition of filters	Sensitivity/recall, precision, number of records retrieved
Whiting 2011 (2010 online)[2]	22 DTA study filters	MEDLINE (Ovid)	MEDLINE (Ovid) search, consulted experts	Details of translations to MEDLINE (Ovid) syntax reported as an appendix	506 references from seven systematic reviews of test accuracy studies that had not used methodological filters in the original search strategy	Compared performance of subject searches with that of filtered searches	Sensitivity/recall, precision, NNR, number of missed records, confidence intervals reported

Study	Filters included	Tested in	Identification of filters	Filter translation	Gold standard	Method of testing	Measures reported
Wong 2006[63]	Three MEDLINE RCT filters; three EMBASE RCT filters; three MEDLINE systematic review filters; four EMBASE systematic review filters	MEDLINE and EMBASE (Ovid)	Strategies developed by the authors and previously published	Not required	Hand-searching of 161 journals for MEDLINE and 55 for EMBASE. Not an external gold standard. RCT records: MEDLINE 930, EMBASE 1256; systematic review records: MEDLINE 753, EMBASE 220	None – reanalysis comparing results of previous publications	Sensitivity/recall, precision, specificity, confidence intervals reported
Studies reporting on the development of one or more filters and their performance in comparison to the performance of previously published filters							
Bachmann 2002[10]	Two DTA study filters, one developed (highest sensitivity × precision) and one published (Haynes 1994[64])	MEDLINE (DataStar)	PubMed Clinical Queries (Haynes 1994[64])	Did not discuss translation or reproduce Haynes[64] strategy used	Hand-search of four journals from 1994 (53 records) and four different journals from 1999 (61 records)	External validation: direct comparison of developed filter and current PubMed filter	Sensitivity/recall, precision, NNR (for developed filter only), confidence intervals reported
Boynton 1998[56]	15 systematic review filters, 11 developed and four published	MEDLINE (Ovid)	Not specified other than published strategies using Ovid Interface	Translation not required	Hand-searching of six journals from 1992 and 1995 – 288 records	Internal validation: compared filter performance against a 'quasi-gold standard'	Sensitivity/recall (described as cumulative), precision (described as cumulative), total articles retrieved, number of relevant articles retrieved

continued

TABLE 9 Review B: table of included studies (*continued*)

Study	Filters included	Tested in	Identification of filters	Filter translation	Gold standard	Method of testing	Measures reported
Devillé 2000[57]	DTA study filters – internal validation: four developed and one published (Haynes 1994[64] sensitive strategy); external validation: one developed (most sensitive) and one published (Haynes 1994[64] sensitive strategy)	MEDLINE (interface unspecified)	Only extensive article on diagnostic filters (Haynes 1994[64])	Not specified but Haynes[64] filter reproduced	Internal validation set: hand-search of nine family medicine journals indexed in MEDLINE (1992–5); database search of MEDLINE (1992–5) to create the 'control set' – 75 records in the gold standard, 137 records in the 'control set'. External validation set: 33 articles on physical diagnostic tests for meniscal lesions; no further details supplied	Internal and external validation: compared retrieval of published and developed strategies	Internal validation: sensitivity/recall, specificity, DOR, confidence intervals reported. External validation: sensitivity/recall, predictive value
Glanville 2006[22]	12 RCT filters, six developed and six published	MEDLINE (Ovid)	Published strategies reporting > 90% sensitivity and with > 100 records in the gold standard used for development	Not specified and filters not reproduced	Database search of MEDLINE (Ovid) (2003) using four clinical MeSH terms. Results assessed to identify indexed and non-indexed trials – 424 records	External validation: compared retrieval in MEDLINE of four clinical MeSH terms with retrieval for each comparator filter	Sensitivity/recall, precision
Haynes 2005[23]	21 RCT filters, two developed (best sensitivity, best specificity) and 19 published	MEDLINE (Ovid)	University filters website and known published articles. Selected strategies that had been tested against gold standards based on a hand-search of published literature and for which MEDLINE records were available from 1990 onwards	Not specified and filters not reproduced	Hand-searching of 161 journals from 2000 – 657 records	External validation: compared performance but full results not presented	Sensitivity/recall, specificity

Study	Filters included	Tested in	Identification of filters	Filter translation	Gold standard	Method of testing	Measures reported
Manríquez 2008[25]	Two RCT filters, one developed and one published (Castro 1999[65])	LILACS	Not specified	Not required (both developed and published filters designed for LILACS)	Hand-searching of 44 journals published between 1981 and 2004 and indexed in LILACS – 267 records	Internal validation: compared ability to retrieve clinical trials included in the gold standard from the LILACS interface	Sensitivity/recall, specificity, precision, confidence intervals reported
Montori 2005[17]	10 systematic review filters, four developed and six published	MEDLINE (Ovid)	'Most popular' published filters	Not specified and filters used not reproduced	Hand-searching of 161 journals indexed in MEDLINE in 2000 – 735 records	External validation: compared filters against validation standard	Sensitivity/recall, precision, specificity, confidence intervals reported
Vincent 2003[14]	Eight DTA study filters, three developed and five published	MEDLINE (Ovid)	Consulted websites, database search of MEDLINE	Not discussed but filters reproduced	References from 16 systematic reviews – 126 records	Internal validation: compared sensitivity of developed and published strategies using reference set of MEDLINE records	Sensitivity/recall
White 2001[19]	Seven systematic review filters, five developed and two published	MEDLINE (Ovid CD-ROM 1995–September 1998)	Not specified	Translated some filters from MEDLINE (Dialog) to MEDLINE (Ovid) syntax	Hand-searching of five journals from 1995 and 1997; quasi-gold standard of systematic reviews – 110 records	Internal validation: compared performance in the 'real-world' search interface using quasi-gold standard	Sensitivity/recall, precision
Wilczynski 2005[15]	Four DTA study filters, two developed (most sensitive, most specific) and two published (most sensitive and most specific)	EMBASE (Ovid)	Not specified	Not discussed but strategies reproduced	Hand-searching of 55 journals from 2000 – 97 records	Internal validation: compared performance of developed and published filters in retrieving 'methodologically sound' diagnostic studies	Sensitivity/recall, precision, specificity, accuracy, confidence intervals for differences between developed and published filters reported

DOR, diagnostic odds ratio; NHS EED, NHS Economic Evaluation Database.
Reproduced with permission from Harbour *et al.*[46] © 2014 The authors. Health Information and Libraries Journal © 2014 Health Libraries Journal. *Health Information & Libraries Journal*, **31**, pp. 176–194.

TABLE 10 Review B: excluded studies

Reference	Reason for exclusion
Bardia 2006[55]	Study compared the performance of filters for complementary and alternative medicine studies rather than RCTs
Kastner 2009[60]	Study examined the performance of the PubMed Clinical Queries sensitive search filter for diagnostic studies in MEDLINE and EMBASE. This was a comparison of a single filter translated to two interfaces and not a comparison of the performance of multiple filters
Royle 2005[62]	Study did not test filters. Study assessed the effectiveness of CENTRAL database methods for the identification of RCTs and the proportion of RCT records that included the term random$

Gold standards

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

Studies included in this review used a range of techniques to identify and/or create a gold or reference standard against which to test the performance of multiple filters. One study did not use a gold standard;[33] instead, each of the filters was combined with single terms describing four topics (hypertension, hepatitis, diabetes and heart failure) and the retrieved studies were checked to confirm whether or not they were RCTs.

The size of the gold or reference standards used to test filter performance ranged from 33 to 1955 records. None of the studies included in this review reported whether or not they had carried out a sample size calculation when developing their gold or reference standard (a sample size calculation is a statistical process that determines the minimum number of records required for a gold standard to provide accurate estimates of performance). Four of the DTA filter studies[2,14,49,58] and one RCT filter study[22] limited their gold standard to specific clinical topics.

Ten studies developed their gold or reference standards by hand-searching journals.[10,15,17,19,23,25,56,57,61,63] The number of journals hand-searched ranged from 4 to 161. The time span covered by hand-searching varied from 1 to 23 years. All of the studies using hand-searching had specific criteria for the identification of the desired study type for inclusion in their gold or reference standard.

Of the 10 studies identifying their gold or reference standard from hand-searching journals, eight were studies in which the authors had developed new search filters and then compared those filters with existing filters.[10,15,17,19,23,25,56,57] One study that created a reference standard from hand-searching journals created a 'control set' of records from the same group of journals that were not of the desired study design.[57]

Five studies developed a gold or reference standard based on the studies included in systematic reviews [relative recall (RR) gold standard][2,14,48,49,58] and four studies used database searches to identify records to include in their gold standard.[22,56,58,59] The number of completed systematic reviews used as a source of gold standard records varied: one study used included studies from 27 systematic reviews,[48] one used included studies from two reviews,[58] one used included studies from seven reviews of DTA studies[2] and a fourth used studies included in a single case study review.[49] One study that developed a DTA study filter and compared it with published filters used the studies included in 16 reviews as the gold standard.[14]

Translation of filters

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

Four of the studies included in this review did not translate or adapt the filters that were being compared because the filters had been developed in the same interface as was used in the performance comparison.[25,33,56,63] When one or more filter required translation, most of the studies comparing the performance of existing filters reported the complete details of the changes made so that the accuracy of the translation could be verified.[2,48,58,59,61] In contrast, most of the studies reporting the development of new filters that included a comparison with existing filters did not mention the requirement to translate any of the filters or provide details of the translation, so it is unclear if valid comparisons were being made.[10,17,22,23,57] The review of economic evaluation filters applied an exclusion strategy (animal studies and publication types such as letters and editorials, which are unlikely to be economic evaluations) to filters being tested in MEDLINE and EMBASE.[59]

Methods of testing

Four of the filter studies that used included studies from systematic reviews as their gold or reference standard replicated the original searches when possible with the addition of the filters being tested.[2,48,49,58] None of the original searches incorporated a study method search filter.[2,48,49,58] A fifth study using references from systematic reviews as a reference standard combined the filters with 'terms for deep vein thrombosis' but did not specify what these terms were or if the original search strategy was used.[14]

The performance analyses carried out by Leeflang *et al.*[48] and Ritchie *et al.*[49] occurred after the original reviews (on which the gold or reference standard was based) had been undertaken and therefore attempted to recreate a 'historical' search. Ritchie *et al.*[49] noted a small discrepancy in the number of records retrieved between the original searches and the rerun searches, whereas Leeflang *et al.*,[48] who could replicate only 6 out of 27 reviews, did not provide details of any differences in the numbers of retrieved records. Using the complete reference standard from the original reviews, Leeflang *et al.*[48] tested whether those studies were captured by the filters being compared.

Two studies did not provide any information about whether the performance analysis had been undertaken concurrently with the reviews or at a later date.[14,58] The review by Whiting *et al.*,[2] which was published online in 2010 and to which we had prepublication access at the time of our study, recreated the original subject search and compared using the subject search alone with using the subject search combined with 22 other filters.

Four studies by the McMaster Hedges team at McMaster University used their internally developed database for testing filters, with the DTA, RCT and systematic review subsets acting as gold standards.[17,23,61,63] One of these studies did not undertake any new analysis but collated the results from previous publications that had used a common gold standard.[63]

The economic filters study identified a gold standard by searching the NHS Economic Evaluation Database (NHS EED).[59] Published MEDLINE and EMBASE economic filters were then tested for their ability to retrieve these gold standard records from MEDLINE and EMBASE. Corrao *et al.*[33] had no gold standard but manually checked whether the records retrieved after applying the filters were RCT studies.

Studies that compared new search filters with existing filters can be divided into two groups based on the type of gold standard used to compare filter performance. One group used a reference standard that had not been used to develop the new filter strategy so that all of the filters in the comparison underwent external validation.[10,17,22,23,57] In other words, the performance of all of the filters being compared was tested in a set of records that had not been used to develop any of the included filters. The other group of studies used the same reference standard that had been used in the development of the new filters, so that, although the new filters underwent only internal validation (filter performance was tested only on the one set of records that had also been used to develop the new filters), the comparison filters underwent external validation.[14,15,19,25,56] The methodology used in the latter group risks introducing bias in favour of the new filters.

Performance measures reported

The most commonly reported performance measures in studies comparing the performance of search filters were sensitivity/recall and precision (*Table 11*). A total of 16 studies reported sensitivity/recall[2,10,14,15,17,19,22,23,25,49,56–59,61,63] and 13 studies reported precision values.[2,10,15,17,19,22,33,49,56,58,59,61,63] Specificity was reported in seven studies.[15,17,23,25,57,61,63]

In one study that did not use a gold standard or reference standard, sensitivity could not be calculated and instead the proportion of retrieved records that met the authors' criteria for being a RCT was reported.[33] In another study the proportions of gold standard records retrieved and missed for each filter were reported.[48] When the original search strategy could not be replicated, this article reported the NNR.[48] Bachmann *et al.*[10] reported the NNR for the filter that they developed but not the previously published filter that they used as a comparator. Whiting *et al.*[2] reported the NNR and the number of records missed from the reference set.

No studies comparing the performance of two or more existing filters reported accuracy values (the number of records correctly retrieved or correctly not retrieved as a proportion of all records). The study by Manríquez[25] reporting the development of a RCT filter for the LILACS database did report accuracy values for

TABLE 11 Review B: measures reported in filter performance comparisons

Performance measure	Study design being identified	Number of studies reporting the measure
Sensitivity/recall	Economic evaluation	1
	DTA study	7
	RCT	5
	Systematic review	3
Precision	Economic evaluation	1
	DTA study	5
	RCT	4
	Systematic review	3
Specificity	Economic evaluation	0
	DTA study	2
	RCT	4
	Systematic review	1
Accuracy	Economic evaluation	0
	DTA study	1
	RCT	1
	Systematic review	0
NNR	Economic evaluation	0
	DTA study	3
	RCT	0
	Systematic review	1
Other (as detailed in text)	Economic evaluation	0
	DTA study	4
	RCT	1
	Systematic review	1

the new filter, as did the study by Wilczynski *et al.*[15] for their newly developed DTA study filters. Additional measures reported in performance comparisons were:

- number of records retrieved[49]
- retrieval gain (absolute and percentage variations in the number of citations retrieved)[33]
- the proportion of articles missed per original review[48]
- the proportion of articles not identified per year[48]
- diagnostic odds ratio (DOR) (the odds of being truly relevant among the relevant divided by the odds of being assessed as relevant among the irrelevant)[57]
- the number of relevant articles retrieved.[56]

Confidence intervals surrounding performance results were reported by three of the studies that compared the performance of existing search filters.[2,61,63] Five of the studies comparing the performance of developed search filters with that of existing search filters reported confidence intervals.[10,15,17,25,57]

Methods used to display performance comparisons/data

All of the studies included in the review displayed the results using a table format, with only two studies supplementing tables of results with graphical (non-tabular) displays of comparative data.[2,48] None of the studies reporting the development of new filters displayed comparative performance in a graphical format.[10,14,15,17,19,22,23,25,56,57]

The majority of tables presenting performance comparison data displayed the filters in rows and performance measures in columns (an example is provided in *Table 12*). The results in the tables in all included studies were provided as percentages or proportions. Within tables, authors generally listed filter results in descending order by the measure of interest, for example decreasing sensitivity. Four studies reporting the development of a filter only included data on comparative performance in the text of the study report.[10,23,25,57]

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

- the databases in which the filters were tested[15,63]
- strategy type (sensitive strategy, specific strategy, optimised strategy)[15,63]
- filter criteria (sensitive, accurate, etc.)[48]
- filter alone compared with a clinical subject strategy[58]
- use or not of an exclusion strategy[59]
- clinical topic considered in the performance testing[33,58]
- subject search alone compared with the same subject search with each test filter[2]
- author or source of published filters[15,17]
- descending order of cumulative precision or cumulative sensitivity.[56]

TABLE 12 Review B: example of a filter performance comparison table as commonly presented in the literature

		Filter	
Filter	Number of records retrieved	Sensitivity (%)	Precision (%)
RCT filter A	*n*	X	Y
RCT filter B	*n*	X	Y
RCT filter C	*n*	X	Y

Reproduced with permission from Harbour *et al.*[46] © 2014 The authors. Health Information and Libraries Journal © 2014 Health Libraries Journal. *Health Information & Libraries Journal*, **31**, pp. 176–194.

Tables were also used to present information on the number of studies retrieved[58] and the specificity, sensitivity and precision of single terms.[63] One study that reported highest precision combined with sensitivity of > 69% showed the results of the filters meeting these criteria in a separate table.[49]

Leeflang *et al.*[48] used a bar graph to display the average proportion of retrieved and missed gold standard records per filter tested (*Figure 1*). Whiting *et al.*[2] presented the overall sensitivity and specificity of each filter tested in a forest plot, including confidence intervals (*Figure 2*).

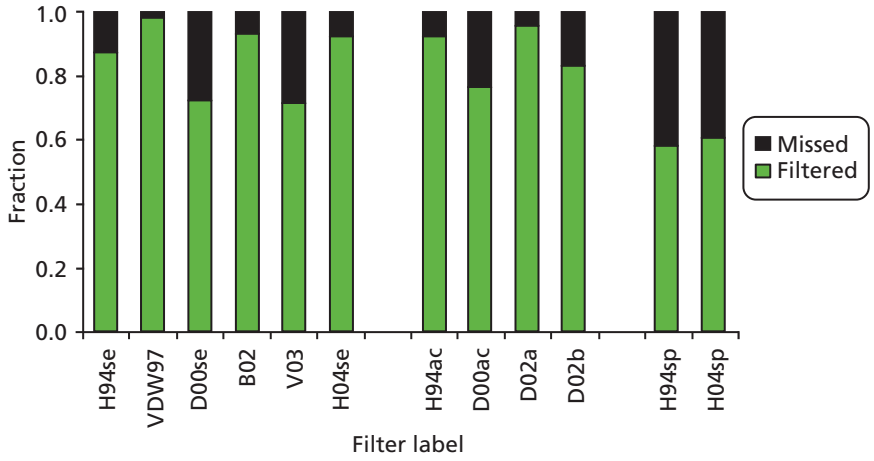

FIGURE 1 Review B: bar chart displaying the comparative performance of filters for DTA studies as published by Leeflang *et al.*[48] Republished with permission of Elsevier from the *Journal of Clinical Epidemiology*, Use of methodological search filters to identify diagnostic accuracy studies can lead to the omission of relevant studies, Leeflang MM, Scholten RJ, Rutjes AW, Reitsma JB, Bossuyt PM, **59**(3), pp. 234–40, copyright 2006;[48] permission conveyed through Copyright Clearance Centre, Inc.

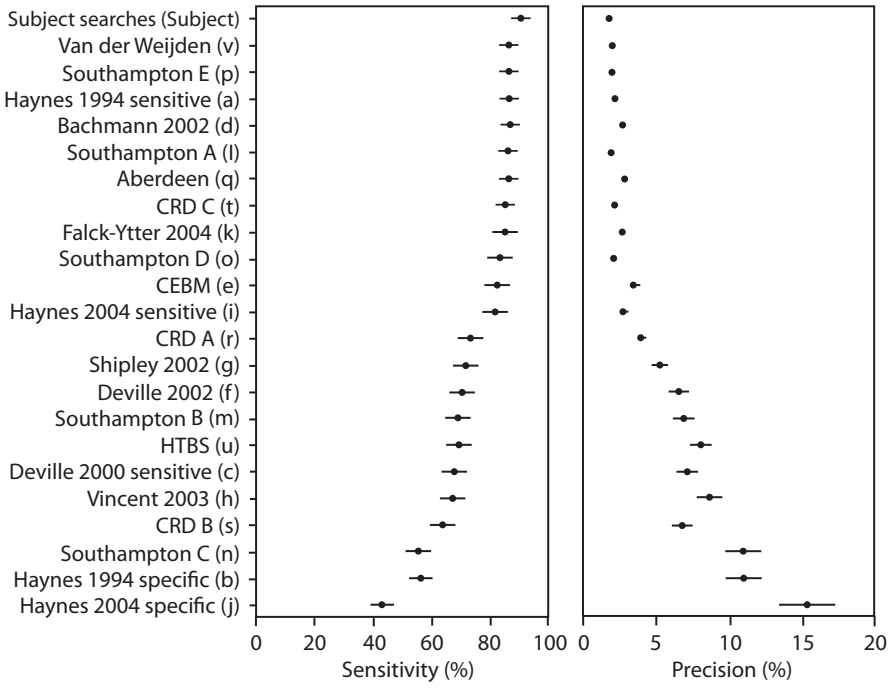

FIGURE 2 Review B: forest plot of overall sensitivity and precision for each filter in the study by Whiting *et al.*[2] CEBM, Centre for Evidence Based Health; CRD, Centre for Reviews and Dissemination; HTBS, Health Technology Board for Scotland. Republished with permission of Elsevier from the *Journal of Clinical Epidemiology*, Inclusion of methodological filters in searches for diagnostic test accuracy studies misses relevant studies, Whiting P, Westwood M, Beynon R, Burke M, Sterne JA, Glanville J, **64**(6), pp. 602–7, copyright 2011;[2] permission conveyed through Copyright Clearance Centre, Inc.

Discussion

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

The number of comparisons of performance varied across study designs. A single study was identified that compared the performance of economic evaluation filters[59] whereas studies reporting on the performance of DTA study and RCT filters were much more common. As there have been, until recently, several specialist economics databases [NHS EED, the Health Economic Evaluations Database (HEED) and the Cost-effectiveness Analysis Registry], it may be that filters for the retrieval of economic evaluation studies have been given a lower research priority than filters for other study designs such as RCTs and DTA studies.

Reporting methods of comparison

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

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

Many filter development studies did not clearly explain how they had identified filters for inclusion in performance testing. Not reporting how filters were identified and whether or not they were developed in the same interface used for testing could have implications for reliability and bias within the studies. If studies do not report how the filters used in comparisons were identified, it is not possible to determine whether the filters were selected in an unbiased fashion or whether they might have been preferentially selected to suit the test environment. In this review, studies reporting the development and testing of one or more filters all found that the new filter performed better than the existing filters used as comparators. This makes it particularly important that studies clearly report how filters are selected and the comparison performed, as otherwise this could be a sign of bias in the results.

Details about the translation of published filters for different interfaces were lacking in many filter development studies. Generally, more details about methods of translation were provided in studies that reported filter performance comparisons separately from the development of new filters. Combined with the lack of information about the original interface used in the development of published filters, the lack of translation details in many filter development studies makes it almost impossible to determine the accuracy of any alterations. As incorrect or imprecise translation of a filter is likely to impact on the results retrieved, the lack of methodological detail provided is a cause for concern.[66]

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

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

Reporting performance measures

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

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

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

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

Limitations of this review

There are a number of potential limitations to this review. It was not possible to undertake a full systematic review because of time constraints. It was also not possible to review all filters for all study methods. The review was, however, focused on study types that were felt to be the key study designs of current interest in evidence-based health research (namely RCTs, DTA studies, systematic reviews or economic evaluation studies). Finally, research carried out on the performance of multiple search filters that has not yet been published or has been presented only at conferences was excluded from the review, possibly resulting in

some alternative formats for the presentation of results being missed. Conference abstracts, however, would be likely to report even fewer methodological details than full articles included in this review.

Key findings

- The main measures of search filter performance reported in the literature are sensitivity/recall, precision and specificity.
- Filter performance comparison studies most commonly report highest sensitivity, highest precision and optimal/balanced filter strategies.
- Articles reporting the development of new search filters and a comparison with existing filters provide limited methodological details.
- Tables are the most frequently used method for reporting the results of filter performance comparisons but graphs may be more useful.

Recommendations

The following recommendations for the presentation of filter performance comparisons are made based on the results of this review.

- Studies that compare search filter performance should explicitly report the methods and results to help searchers identify the most appropriate filter for their particular purpose.
- Studies presenting the development of new search filters that include comparisons with existing filters should present detailed methods describing how the performance comparisons were undertaken.
- One or more gold or reference standards should be used for testing filter performance.
- Search filters should be validated on gold or reference standards that are different from those from which they were developed.
- The size of the gold or reference standard(s) should be clearly stated and a sample size calculation presented to justify the size of the standard(s).
- Any translation of filters should be specifically reported in all articles in which a filter has been used in a different interface from that in which it was developed.
- Results should be presented systematically, identifying clearly the best-performing filter for specific purposes (sensitive strategy, specific strategy, balanced strategy).
- When tables of performance results are provided, a consistent format and order should be used to make the information easy to extract.

Measuring performance in diagnostic test accuracy studies (review C)

Introduction

Performance measurement of search filters can be seen as analogous to DTA in that DTA studies aim to reliably differentiate those with a specific disease (relevant studies for searchers) from those who do not have the disease (irrelevant studies for searchers). They also aim to be as accurate as possible in distinguishing cases of disease from cases of non-disease, by minimising false positives (positive results for those who do not have the disease) and false negatives (missing cases of people with a disease). Similarly, search filters aim to identify all relevant studies (true positives) while aiming to minimise the retrieval of irrelevant studies (false positives).

This review explores published guidance and recommendations that inform best practice in the measurement and reporting of DTA and assesses their applicability to the area of search filter performance.

Objectives

- To identify recommended methods for conducting DTA studies and evaluating test performance.
- To identify the diagnostic test performance measurements that have been reported and presented.

- To identify methods to compare DTA performance from primary studies.
- To assess how applicable these measures and methods are to search filter performance and how these measures might add value to the filter selection process.

Methods

We undertook literature searches of electronic databases to identify articles that reviewed methodological aspects of undertaking DTA studies and DTA reviews or provided guidelines and other recommendations on how DTA studies or reviews should be carried out and how the results should be reported. These searches were supplemented by consulting key HTA agencies and Cochrane websites for relevant reports or recommendations.

The following databases were searched in October 2011: Cochrane Methodology Register, The Cochrane Library (Issue 4, 2011), Medion (October 2011), MEDLINE (1950 to October Week 3 2011), MEDLINE In-Process & Other Non-Indexed Citations (28 October 2011) and EMBASE (1980 to Week 43 2011). Full details of the strategies used are reproduced in *Appendix 2* along with a list of websites that provided potentially useful reports.

From the electronic database searches, 1454 records were retrieved, which was reduced to 972 records after deduplication. After screening titles and abstracts, 97 records were selected as being potentially useful (*Figure 3*). The full articles were obtained and read for relevance. In addition, eight reports were obtained from organisation websites. Forty-seven of these reports contributed information to the review.[36,67–112] A list of the remaining 58 retrieved documents that were excluded from the review is provided in *Appendix 3*. Studies were excluded because they were considered to be irrelevant, described issues or methods that were better expressed or more thoroughly considered in another publication or were duplicate publications. A flow chart showing the selection process for inclusion of studies in the review is provided in *Figure 3*.

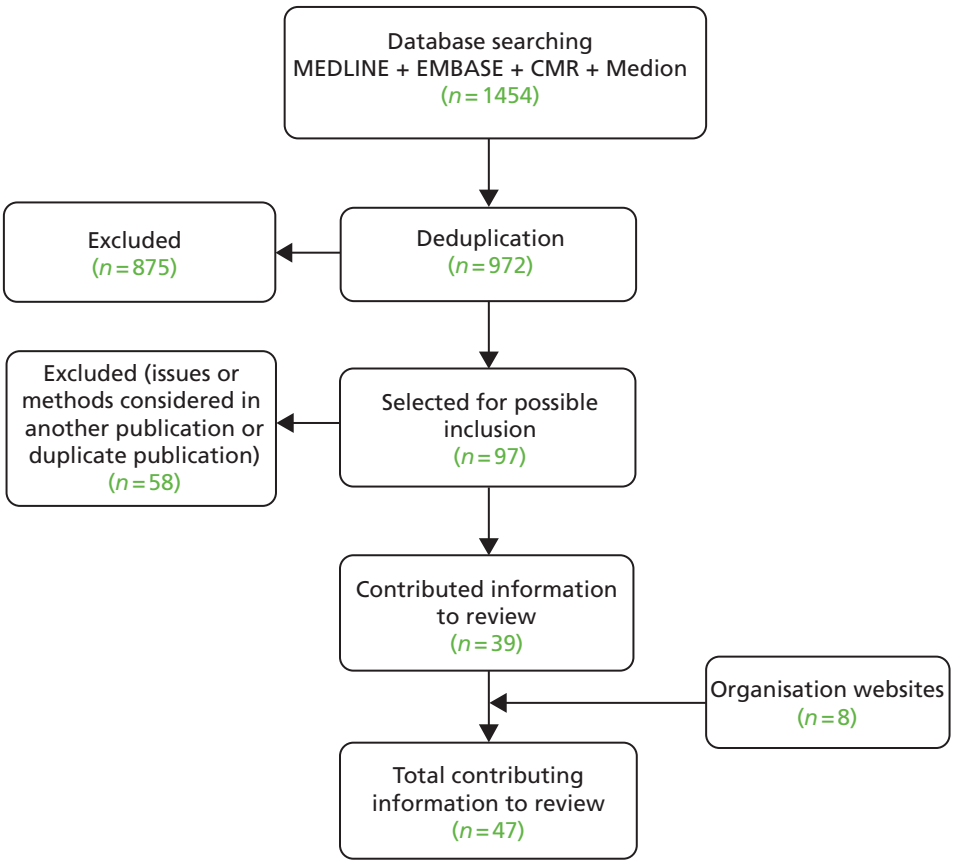

FIGURE 3 Review C: selection of reports for inclusion in the review. CMR, Cochrane Methodology Register.

We acknowledge that there has been a regrettable delay between carrying out the project, including the searches, and the publication of this report, because of serious illness of the principal investigator. The searches were carried out in 2010/11.

Results for diagnostic test accuracy studies

Conducting diagnostic test accuracy studies

Diagnostic test accuracy measures the ability of the diagnostic test being evaluated, the index test, to distinguish between patients with and patients without the targeted disease or condition.[67] The results are verified against the results of a reference standard in the same group of patients. The reference standard is independent of the index test and is usually the best available method to identify patients with the target condition.[68,69] When a comparator test is also under evaluation, the index and comparator test must be evaluated against the same reference standard and in the same population.[69] In the absence of a suitable reference standard a number of alternative methods have been proposed.[70–72]

Test accuracy is not fixed and can vary between patient subgroups, with disease severity, in different clinical settings and with different test interpreters.[67] Several guidance documents describe how these variations in the design and conduct of diagnostic tests can lead to bias, resulting in substantial differences being observed between primary studies.[69,73–76] The effects of different types of bias have been estimated using empirical data.[76–79]

As diagnostic tests do perform differently in different populations, the importance of testing in a suitable sample of patients receives much attention in the literature. The patient sample should be representative in terms of the disease severity of the target population for whom the test is intended, to avoid spectrum bias (i.e. the variation in the sensitivity and/or specificity of a diagnostic test when applied to people of different ages, genders, nationalities or specific disease manifestations).[69,73,75,80] Ideally, patients should be recruited consecutively or randomly in a single cohort and be unselected by disease state.[74] Case–control studies are likely to lead to bias because patients with and without the condition are recruited using different sets of criteria[69,73] and because they overestimate diagnostic accuracy.[77] Other main sources of bias relate to the unsuitability of the reference standard, how the reference and index tests have been undertaken, interpreter blinding and interpretation of the results.[79]

Uncertainty around estimates of diagnostic accuracy decreases with increasing sample size[75] and it is recommended that sample size calculations should be undertaken during study planning to ensure that a reasonably precise estimate of test accuracy can be achieved.[81,82] Tables have been published to assist in determining the minimum sample size required[83] for a DTA study once the prevalence of the target condition in the population as well as the expected sensitivity have been determined. However, two reviews of DTA studies found that very few studies gave any consideration to sample size.[81,82]

The Quality Assessment of Diagnostic Accuracy Studies (QUADAS) tool has been developed to assist researchers to assess the quality of primary DTA studies[84,85] and as such provides a useful guide to the issues that should be addressed when undertaking a DTA study. The questions cover aspects of methodology that are thought to make a difference to the reliability of a study, such as the suitability of the patient sample and the reproducibility of the reference standard and index test. Poor reporting of DTA studies, however, can make applying the QUADAS tool difficult.[78] Since the searches for this review were undertaken, a revised version of the QUADAS tool has been published. The QUADAS 2 tool, which is applied in four phases, will, according to the publishers, allow for a more transparent rating of bias and the applicability of primary diagnostic accuracy studies than the original QUADAS tool.

Measuring diagnostic test accuracy

Contingency table

The primary outcomes of interest in DTA studies are the data required to populate 2 × 2 contingency tables presenting the presence or absence of the target condition or disease, as defined by the reference standard against the result of the index test (*Table 13*). From this all DTA measures can be derived.

Measures

Table 14 describes the measures of diagnostic accuracy that are commonly calculated, namely sensitivity, specificity, likelihood ratio (LR), DOR and predictive value.

Two statistical measures of diagnostic accuracy are traditionally used in a clinical setting: the true positive rate or the sensitivity of the test (the proportion of those with the disease who have an abnormal test result) and the specificity of the test (the proportion of those without the disease who have a normal test result). To rule out a diagnosis a test must have high sensitivity whereas to confirm a diagnosis a test must have high specificity.[69,73,80] Both measures are susceptible to spectrum bias[76,86] but are not directly influenced by prevalence.[76]

The predictive value is the probability of the test correctly diagnosing patients. The PPV is the proportion of patients with a positive test result who are correctly diagnosed. Conversely, the negative predictive value (NPV) is the proportion of patients with a negative test result who are correctly diagnosed. Predictive values depend on the prevalence of the condition in the population being tested. When prevalence is high, it is more likely that a positive test result is correct and a negative result is wrong.[86,87]

TABLE 13 Review C: contingency table

	Disease		
Test result	Present	Absent	Total
Positive	A (true positive)	B (false positive)	A + B (test positive)
Negative	C (false negative)	D (true negative)	C + D (test negative)
Total	A + C (disease)	B + D (no disease)	A + B + C + D

TABLE 14 Review C: measures of diagnostic accuracy

Measurement	Formula	Definition
Sensitivity	A/(A + C)	Proportion of patients with the disease correctly identified by the test
Specificity	D/(D + B)	Proportion of patients without the disease correctly identified by the test
LR	LR for positive result (LR+) = [A/(A + C)]/[B/(B + D)] LR for negative result (LR–) = [C/(A + C)]/[D/(B + D)]	How many times a person with the disease is more likely to receive a particular test result (positive or negative) than a person without the disease
DOR	[(A/C)/(B/D)] = (AD/BC)	Summary measure of the diagnostic accuracy of a diagnostic test
Predictive value	PPV = A/(A + B)	Proportion of patients with a positive test result who are correctly diagnosed
	NPV = D/(C + D)	Proportion of patients with a negative test result who are correctly diagnosed

LR–, negative likelihood ratio; NPV, negative predictive value.

Likelihood ratios describe the performance of diagnostic tests and can be useful in a clinical setting. The ratio describes whether or not a test result usefully changes the probability that a condition exists. The LR+ is the probability of a person who has the disease testing positive divided by the probability of a person who does not have the disease testing positive. A LR+ of > 10 and a negative likelihood ratio (LR–) of < 0.1 are judged to provide convincing diagnostic evidence.[88] Their interpretation, however, depends on the clinical context.[87]

The DOR is a summary measure of the diagnostic accuracy of a diagnostic test. It is calculated as the odds of positivity among diseased persons divided by the odds of positivity among non-diseased persons. When a test provides no diagnostic evidence then the DOR is 1.0.[89] This measure has a number of limitations. In particular, it combines sensitivity and specificity into a single value, hence losing the relative values of the two, and is difficult to interpret clinically.[87]

Sensitivity and specificity are based on binary classification of test results (either positive or negative). Test measures, however, are often categorical or continuous and so a cut-off point must be defined to classify results as either positive or negative. As the threshold shifts, the sensitivity and specificity of a test will change, with an increase in one resulting in a decrease in the other. This trade-off at different thresholds can be presented graphically in a receiver operating characteristic (ROC) curve, describing the relationship between the true-positive value (sensitivity) and the false-positive value (1 – specificity), and can be used to identify a suitable threshold for clinical practice.[69] *Figure 4* displays a sample ROC curve of test performance using different threshold values from ≥ 5 to > 25.

The Q* value is the point on the ROC curve where sensitivity equals specificity and can be used as a single indicator of overall test performance when there is no preference for maximising sensitivity (minimising false negatives) or specificity (minimising false positives) but can give misleading results if used to compare performance between tests.[69,90] Overall, diagnostic accuracy is summarised by the area under the curve (AUC) and ranges from 0.5 (very poor test accuracy and equivalent to chance) to 1.0.[69,87] The more accurate the test, the more closely the curve approaches the top left hand corner and has a value close to 1.0.

Whiting *et al.*[87] have undertaken an overview of the various types of graphical presentations that have been used in the DTA literature and describe other graphical displays that could be used to present DTA data. These include dot plots, box-and-whisker plots and flow charts (*Figure 5*).

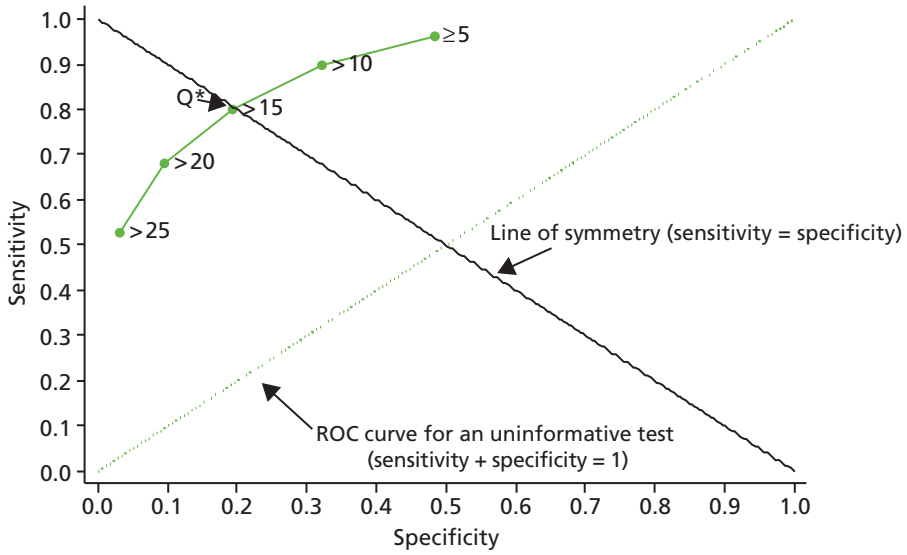

FIGURE 4 Review C: example ROC curve.

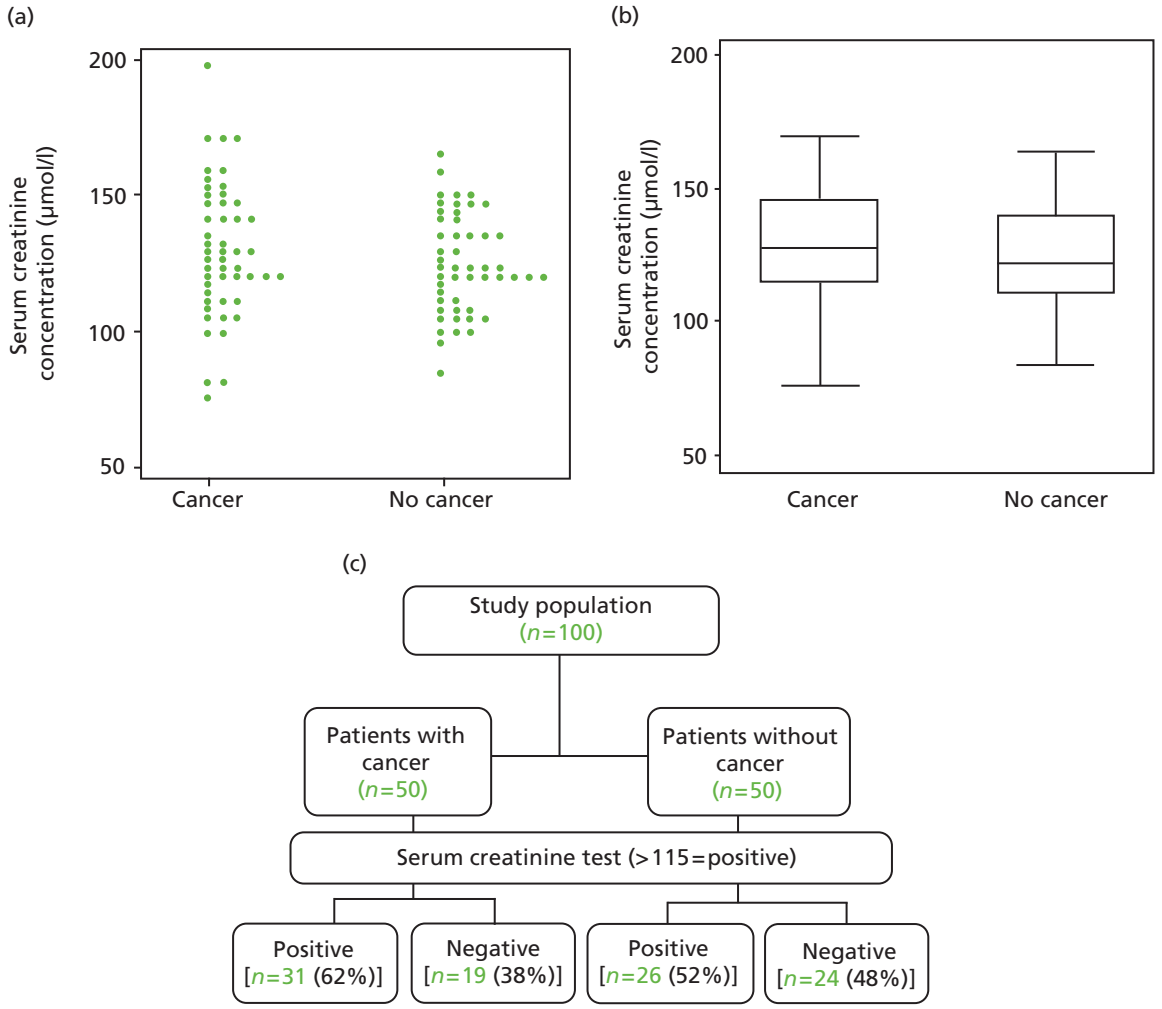

FIGURE 5 Review C: example graphical displays for primary study data. (a) Dot plot; (b) box-and-whisker plot; (c) flow chart.

Dot plots are used for test results that take many values and display the distribution of results in patients with and without the target condition but do not directly display diagnostic performance. Box-and-whisker plots summarise the distributions of true-positive and true-negative groups by a continuous measure. Flow diagrams depict the flow of patients through the study, for example how many patients were eligible, how many entered the study, how many of these had the target condition and the numbers testing positive and negative.

Reporting of test accuracy results

The Standards for the Reporting of Diagnostic Accuracy Studies (STARD) statement[68] provides guidance on how DTA studies should be reported to provide transparency and allow the reader to assess the validity of a study. Full details on participants, method of recruitment, reference and index tests, statistical methods and results are required. Several predominantly small reviews of between 16 and 243 studies[91–98] have looked at the reporting of DTA studies and found poor description of the methods used. Studies either lacked completeness of reporting, with < 50% of studies reporting over half of the STARD items,[95,96] or lacked clarity, hence making assessment difficult.[97] These reviews concluded that the STARD statement seems to have resulted in little improvement in study reporting. Most of these reviews, however, included studies that were published prior to or soon after the STARD statement was published[91,92,98,99] and so it may be the case that insufficient time had elapsed to make a valid assessment.

Guidance documents provide few recommendations about which DTA measures should be reported. The choice of accuracy measures presented depends on the aims of a particular study and on who is likely to use

the information. For example, LRs may be more useful in a clinical setting as they can be used to calculate the probability of disease for individual patients, whereas DORs are difficult to interpret clinically. US,[75] Australian[76] and UK[69] guidance suggests that the 2 × 2 contingency table together with sensitivity and specificity pairs and LR pairs should be presented, along with 95% confidence intervals.[75,76] The US Food and Drug Administration (FDA) also recommends that measures are reported both as fractions and as percentages.[75]

There is some information about measures reported in the literature.[100] In a review of 90 DTA reviews,[101] sensitivity or specificity was the most common measure used to report the results of primary studies (in 72% of reviews); predictive values were included in 28% of reviews; and LRs were included in 22% of reviews. In reviewing the reporting of DTA measures in primary studies, two studies have noted that sensitivity and specificity were reported in most studies, with ROC curves reported in less than half of the studies.[95,96]

There is some evidence that studies rarely present diagnostic information graphically.[87,91,102] In a review of 57 primary studies,[99] 57% used graphical displays to present results. Dot plots or box-and-whisker plots were the most commonly used graphs in the primary studies (in 39% of studies) whereas ROC curves were displayed in 26% of studies.

Methods to compare and synthesise diagnostic test accuracy performance from primary studies

Several HTA organisations, in guidance for undertaking DTA evidence synthesis,[69,76,90,103,104] recommend using the QUADAS tool or a modified version to assess the methodological quality of primary studies. Undertaking a formal assessment provides an indication of the degree to which the included studies are prone to bias[100,102,105,106] and hence the reliability of the study results. A report from the Agency for Healthcare Research and Quality (AHRQ)[100] found that there had been a trend in recent years for an increasing number of DTA reviews to formally assess study quality.

Several organisations have developed guidance on carrying out systematic reviews of DTA studies[69,75,76,90,102–104] and agree that analysis is more complex than for clinical effectiveness. Combining results from individual studies can be problematic because of the methodological variability (heterogeneity) found across the studies. In particular, combining test accuracy studies with heterogeneity can produce biased, and hence inaccurate, results.[74,79,104,107,108]

It is recognised that variability among studies is to be expected. Some of the variability is due to chance, because many diagnostic studies have small sample sizes. The remaining heterogeneity may be the result of differences in study populations or differences in study methods or the result of variation in the diagnostic threshold adopted.[74] Several methods have been described to measure heterogeneity, using graphical plots and statistical tests.[36,76,109] Although it is recommended that such a thorough investigation be undertaken prior to meta-analysis,[69,75,76,86,90,100,102–104] this is often not carried out. In a review of 189 systematic reviews,[109] only 32% investigated heterogeneity and the authors concluded that this underuse reflected uncertainty about the correct approach to adopt.

It is recommended that only studies using the same reference standard, including substantially similar patients and showing minimal heterogeneity should be synthesised by meta-analysis.[69,74,76,90,104] When this type of complex analysis is undertaken it has been recommended that reviewers should enlist the specialist support of an experienced statistician in the field.[36,69,109] When it is not suitable to undertake meta-analysis a narrative approach should be adopted using graphical presentations, such as forest plots and ROC space plots,[69] to provide a visual overview of the results from the included studies.

Paired forest plots (*Figure 6*) can show the spread of estimated values for sensitivity and specificity for each study. Point estimates are shown as dots or squares and can be sized according to the precision of the estimate or sample size. Confidence intervals around the estimate are shown by horizontal lines either side of the point estimate. If meta-analysis is then undertaken, the pooled estimate is displayed as a diamond.

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)
Cheng 2000	57	39	7	72	0.89 (0.79 to 0.95)	0.65 (0.55 to 0.74)
De Dominicis 2001	45	47	7	80	0.87 (0.74 to 0.94)	0.63 (0.54 to 0.71)
Ehsan	59	32	1	59	0.98 (0.91 to 1.00)	0.65 (0.54 to 0.75)
Fillbeck_a 1999	119	145	5	78	0.96 (0.91 to 0.99)	0.35 (0.29 to 0.42)
Grimbergen 2003	378	270	12	257	0.97 (0.95 to 0.98)	0.49 (0.44 to 0.53)
Hendricksen	88	52	6	71	0.94 (0.87 to 0.98)	0.58 (0.48 to 0.67)
Hungerhuber 2007	1484	1339	129	1678	0.92 (0.91 to 0.93)	0.56 (0.54 to 0.57)
Jeon 2001	145	72	3	54	0.98 (0.94 to 1.00)	0.43 (0.34 to 0.52)
Jichlinski 1997	97	46	12	60	0.89 (0.82 to 0.94)	0.57 (0.47 to 0.66)
Jichlinski 2003	108	57	35	221	0.76 (0.68 to 0.82)	0.79 (0.74 to 0.84)
Kreigmair 1996	126	110	2	195	0.98 (0.94 to 1.00)	0.64 (0.58 to 0.69)
Sim 2005	61	10	13	95	0.82 (0.72 to 0.90)	0.90 (0.83 to 0.95)
Witjes 2005	23	0	4	1	0.85 (0.66 to 0.96)	1.00 (0.03 to 1.00)
Zumbraegel 2003	125	186	8	89	0.94 (0.88 to 0.97)	0.32 (0.27 to 0.38)

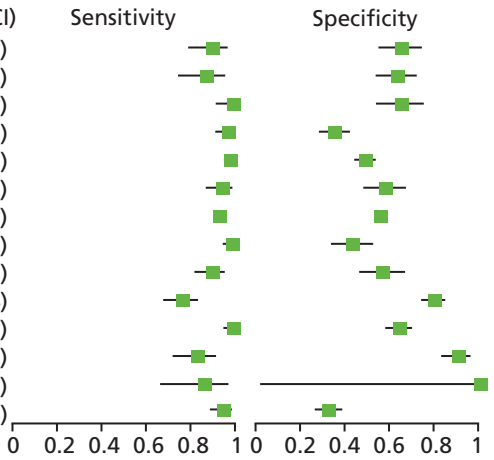

FIGURE 6 Review C: example of a paired forest plot. FN, false negative; FP, false positive; TN, true negative; TP, true positive.

ROC space plots (*Figure 7*) present the relationship between sensitivity and specificity, with each point representing the summary performance for each study.[69]

When performance measures are pooled, separate meta-analyses of sensitivity and specificity data are both the simplest and the most useful approach.[69,104] Such an approach, however, assumes that all included studies are using the same threshold value. Summary ROC (SROC) curves are a form of meta-analysis in which the result is a ROC curve with each data point representing the paired estimate of sensitivity and 1 – specificity from the separate studies (*Figure 8*). Hierarchical and bivariate statistical models have been developed to estimate the SROC curve.[110,111] The SROC curve is a useful presentation when a threshold effect is observed. The curve provides a global summary of test accuracy and, as with a ROC curve, shows the trade-off between sensitivity and specificity at different threshold levels. It does not, however, provide a single statistic of overall test performance[104] and a review has indicated slow uptake of these newer methods.[112]

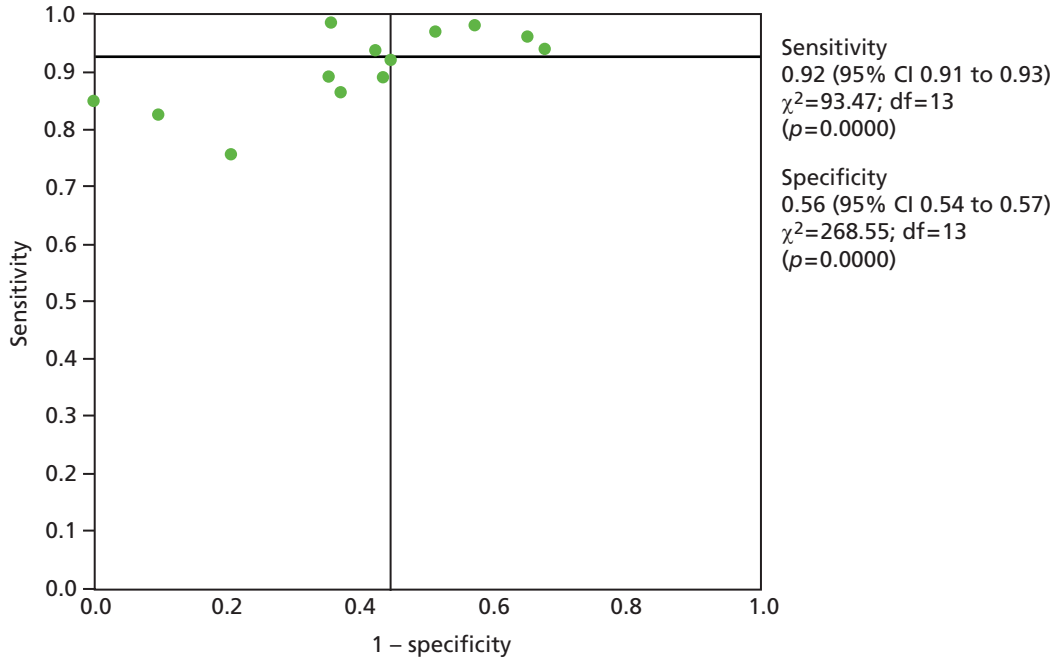

FIGURE 7 Review C: example of a ROC space plot showing summary sensitivity and specificity. df, degrees of freedom.

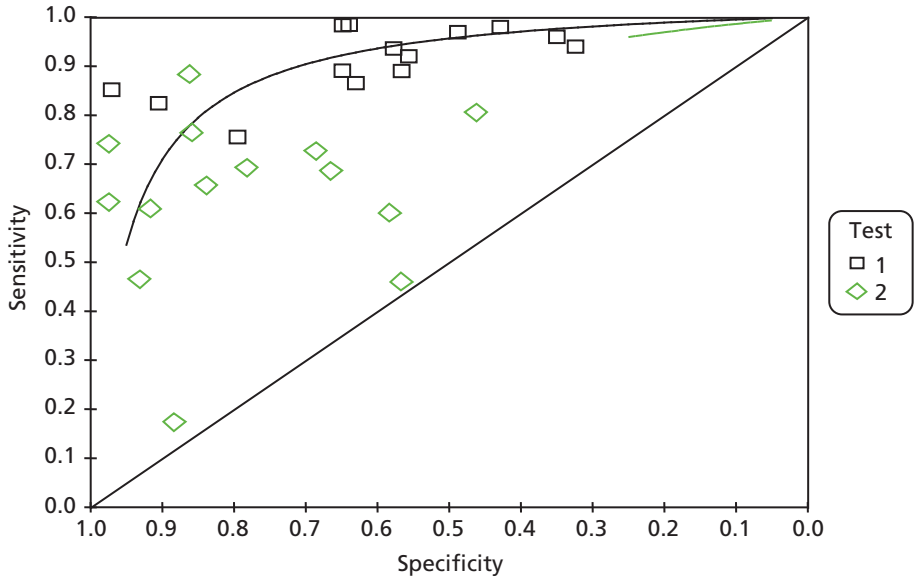

FIGURE 8 Review C: example of a paired SROC curve, comparing the accuracy of test 1 with that of test 2.

Other graphical methods that can be used to present data in a way that is useful in a clinical context have been suggested.[87] The two main methods are LR nomograms and the probability-modifying plot. These graphs enable the clinician to estimate the post-test probability of a patient having the disease, based on their pretest probability, when the LRs of tests are known.

Whiting *et al.*[87] reviewed the graphical presentation of diagnostic information in 49 systematic reviews. Just over half (53%) of the reviews used graphical displays to present the results. ROC plots were the most common type of graph and were included in 22 reviews (45%), whereas forest plots were used in 10 reviews (20%) to display individual study results. In another review of DTA reviews, Honest and Khan[101] found that, when meta-analysis had been undertaken, pooled sensitivity or specificity was reported in 35 out of 60 (58%) reviews, pooled predictive values in 11 out of 60 (18%) reviews, pooled LRs in 13 out of 60 (22%) reviews and pooled DORs in five out of 60 (8%) reviews. SROC plots were reported in 44 out of 60 (73%) of the meta-analyses. Dinnes *et al.*[109] noted that, out of 189 systematic reviews included in their review, 30% had involved narrative analysis and, when meta-analysis had been undertaken, 52% statistically pooled data, 18% reported SROC plots and a further 30% employed both techniques.

Summary

- Diagnostic test accuracy studies should be carried out on a sample of patients who are representative of the target population, particularly in terms of disease state, and should use an appropriate reference standard with interpreter blinding to previous test results.
- Sensitivity (true-positive rate) and specificity (true-negative rate) are the most commonly reported outcomes and are subject to spectrum bias.
- Predictive values, used to calculate the probability of a test giving a correct result, are influenced by the disease prevalence in the population.
- LRs are useful in a clinical setting to determine the probability of a patient having the target disease.
- DORs provide a summary measure combining sensitivity and specificity but are difficult to interpret clinically.
- ROC curves present sensitivity and specificity pairs at different test thresholds, whereas the AUC gives an overall value of DTA.
- International HTA organisations that have addressed the issue recommend that DTA studies should present 2 × 2 contingency tables, sensitivity and specificity pairs and LR pairs.
- Several types of graphical presentations can be used to display DTA data but these have not been used extensively in the DTA literature.
- In undertaking systematic reviews of DTA studies, heterogeneity between studies is a common feature and should be investigated before combining data in a meta-analysis.
- A narrative approach, presenting forest plots and ROC space plots, is recommended when heterogeneity exists.
- Poor quality in relation to methodology and reporting affects the inferences that can be drawn from DTA studies.

Applicability to research in search filter performance

Diagnostic test accuracy and search filter studies share similar characteristics in that both evaluate the performance of an index test (or search filter) against that of a reference standard in the same sample of patients (or records). In the clinical literature, the reference standard should be the best available method to identify the 'target condition'. In the search filter literature the reference standard usually refers not to the method per se but rather the set of relevant records that the method has been designed to identify.[51,76] Typically, the reference standard is described as the records obtained by hand-searching a set of journals over a specified time period (i.e. the 'positive' records in the sample to be tested) rather than describing the reference standard as the method used (i.e. 'hand-searching'). Other reference standards used, such as the records of included studies from systematic reviews or studies held in a specialised register, again conflate the method and the sample. In these cases, the method used is implicit: searching and screening to identify relevant studies. Although the terminology is different, the principle is the same: the results of applying the index test or filter to a sample are compared with the results of a method that is considered to be robust.

Methods for conducting a search filter performance study

Guidance on measuring DTA performance emphasises the importance of using a sample of patients who are representative of the intended population, particularly in relation to the target condition, otherwise the study may be subject to spectrum bias. Likewise, when measuring search filter performance of a filter intended for a particular bibliographic database, the set of records on which the filter is tested should be representative of that database.

When hand-searching is undertaken, the selection of journals used should be representative of the journals that are indexed in the bibliographic database for which the filter is intended. In terms of subject/clinical focus this can be problematic because hand-searching is labour intensive and so the requirement to include a representative selection of journals has to be balanced against the need to obtain a sufficient yield of articles efficiently by using specialist high-yield journals. For example, when testing or developing a DTA study filter, hand-searching radiology journals may be an efficient way to provide a good yield of DTA studies but these will not be representative of health-care journals in general. The underlying prevalence in the test sample is likely to be much higher than for the whole database and will result in overestimation of the internal precision of the resulting filter. Other factors to consider in selecting journals might include language (including UK/US variations), impact factors and the inclusion of abstracts in the database records.

Using included studies from reviews or a study register such as CENTRAL is likely to provide a wider range of publication sources. The original search strategies used in the reviews should be sensitive and ideally not include methodological search filters so that bias is not introduced by limitations in the searches. However, the inclusion criteria used to select the studies for the reviews or registers may also introduce bias. For example, the reviews may include only large RCTs so the reference standard under-reports all RCTs on the review topic retrieved by the subject search. This will impact on the measurement of the performance of the search filter, particularly in terms of reducing precision. Reduction in the NNR, which calculates a reduction in the number of records to be screened, may be a more appropriate parameter in these circumstances.

As bibliographic databases have changed over time in terms of both content and indexing vocabulary, the publication span for hand-searched journals and included studies also deserves attention to ensure representative coverage.

The DTA literature mentions sample size as another important issue, although the literature suggests that this is seldom formally reported. This is also the case for search filter performance literature. The performance measures calculated for the test sample are an estimate of the population value and uncertainty around these performance measures (as demonstrated by the confidence intervals) decreases with an increase in the sample size.

Tables have been published to assist in sample size calculations for DTA studies and would be appropriate to use for search filter studies.[83] An example is shown in *Table 15*.

When the prevalence of relevant records across the results set is expected to be < 0.50 (which would be the case in search filter design studies), the following steps can be followed to calculate the sample size:

Reference set:

- for example, based on the assumption that the expected specificity of the filter will be 90% (see *Table 15*, seventh row) and
- if we specify that the minimal acceptable lower confidence limit is, for example, 0.75 (see *Table 15*, sixth column)
- then the minimal sample size for the reference set (*N*cases) is read from the table as 70 records.

TABLE 15 Review C: calculating sample sizes for search filter design studies. Number of cases (and controls) for expected sensitivities (or specificities) ranging from 0.60 to 0.95. Reprinted from the *Journal of Clinical Epidemiology*, Vol 58, Flahault A, Cadilhac M, Thomas G, Sample size calculation should be performed for design accuracy in diagnostic test studies, pp. 859–62, copyright (2005), with permission from Elsevier[85]

Expected sensitivity (or specificity)	Minimal acceptable lower confidence limit								
	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90
0.60	268	1058							
0.65	119	262	1018						
0.70	67	114	248	960					
0.75	42	62	107	230	869				
0.80	28	40	60	98	204	756			
0.85	18	26	33	52	85	176	624		
0.90	13	18	24	31	41	70	235	474	
0.95	11	12	14	16	24	34	50	93	298

Results set:

- the minimum results set is calculated from the equation (*N*cases) + *N*controls, where *N*controls = *N*cases [(1 – prevalence)/prevalence]
- if we assume that the expected prevalence of relevant records is 5% of the hand-search or search results then the results set is calculated as 70 + 70[(1 – 0.05)/0.05] = 70 + 1330 = 1400 records

A lower assumed prevalence would increase the size of the required results set. For example, for a 1% assumed prevalence, the reference set should be 7000 records.

Other main sources of bias mentioned in the DTA literature relate to the suitability of the reference standard (appropriate to the target condition and independent from the index test) and to the methods used in carrying out the test (interpreter blinding and standard interpretation of the results). In terms of search filter testing, there are factors that might affect the independence between the index test and the reference test. For example, when screening journal abstracts, hand-searchers should be unaware of the indexed terms used in the corresponding database records and, when the included studies in a review are used as the reference set, the original search strategy terms should not include any of the search terms being tested. Ideally, the review's search strategies should have no methodological terms.

Irrespective of how the reference standard is obtained, methods should be standardised to help limit variability. When multiple hand-searchers are involved in creating the reference standard, they should work to the same inclusion and exclusion criteria, which match the study type(s) that the test filter is intended to retrieve, and reviewers' reliability should be formally assessed before commencement.

Checklists similar to the QUADAS tool[85] and the STARD statement,[68] but designed for search filter studies, would enable a formal assessment of study quality and might assist search filter researchers to adopt a more consistent and high-quality methodology. Examples of checklists for search filter studies have been reported,[3,4,51] with only that of Bak *et al.*[3] including a scoring system.

Search filter performance measures

In DTA performance measurement, sensitivity and specificity are the most commonly reported values and are judged to be essential by most guidance. Other measures that tend to be reported are PPVs and NPVs, LRs and DORs. For search filter performance, sensitivity (or recall) is almost universally reported, with specificity and precision (equivalent to PPV) the next most frequently reported measures (see reviews A and B).

Specificity and precision (or PPV) are both measures of the false-positive rate; the former is measured in relation to the total number of negatives whereas the latter relates to the number selected by the filter or test.

In situations in which data are highly skewed, as is the case with literature retrieval, when typically a very small fraction of records in a bibliographic database are relevant (positive), precision rather than specificity better captures changes in the false-positive rate. This is because the number of false positives is being compared with a relatively small number of true positives rather than the much larger number of true negatives.[113]

This phenomenon is illustrated by the precision and specificity of the three filters shown in *Table 16*. Filter A has 83% sensitivity, 25% precision and 92% specificity. For filters B and C, the number of relevant records retrieved is the same and so sensitivity is maintained at 83%. The number of retrieved irrelevant records, however, varies. For filter B, the number has more than doubled from 750 to 1750 and consequently precision has been halved to 12.5% whereas specificity has been reduced from 92% to 82%, a reduction of only 11%. A large increase in the number of irrelevant records retrieved has led to a substantial change in precision but a relatively small change in specificity. For filter C, the number of retrieved irrelevant records has increased almost seven-fold, resulting in specificity being reduced by half to 46%. The resulting change in precision of approximately 80%, from 25% to 4.6%, again better reflects the huge increase in number of irrelevant records being retrieved.

In the context of evidence synthesis, a searcher's primary interest is to know how many relevant records have been missed by the search as well as how many retrieved records are irrelevant but will still require to be screened. These factors affect how efficiently and accurately data gathering for evidence synthesis will be carried out. Sensitivity and precision are therefore of most interest. A busy clinician, however, may prefer to retrieve a small set of records in which a high proportion are relevant, and so high precision is very important whereas sensitivity is less important. Knowing the proportion of irrelevant records in a bibliographic database that have not been retrieved, as measured by specificity, is of lesser value.

Likelihood ratios, although useful in a clinical situation for indicating a patient's probability of truly having the target condition, are probably of less use in literature searching because searchers are less interested in individual records. The DOR, sometimes referred to as 'accuracy', is a single indicator of diagnostic performance and has occasionally been calculated in search filter literature. As with a clinical situation, however, it provides a summary measure and hence does not provide as much useful information on performance as other measures.

Presentation of results

In search filter performance studies, tabular presentation of the results is the norm. DTA study guidance suggests several different graphical presentations that can be used, although they seem to be underused in the DTA literature.

In clinical situations, test measurements are frequently continuous in nature and so thresholds are set to define positive and negative results. The trade-off between sensitivity and specificity at different thresholds

TABLE 16 Review C: precision and specificity illustration

Filter	Filter performance	Retrieval	Relevant	Not relevant	Total
A	Sensitivity 83%; precision 25%; specificity 92%	Retrieved	250	750	1000
		Not retrieved	50	8950	9000
		Total	300	9700	10,000
B	Sensitivity 83%; precision 12.5%; specificity 82%	Retrieved	250	1750	2000
		Not retrieved	50	7950	8000
		Total	300	9700	10,000
C	Sensitivity 83%; precision 4.6%; specificity 46%	Retrieved	250	5238	5488
		Not retrieved	50	4462	4512
		Total	300	9700	10,000

is often graphically presented in a ROC plot. This situation does not occur in standard literature searching: a search filter produces a binary result, either selected or not. At the filter development stage, however, a ROC plot could be a useful way to display the performance characteristics of variations in a filter, showing the change that results from the inclusion or exclusion of particular search terms.

Other graphical presentations that have been used in the DTA literature include dot plots, box-and-whisker plots and flow diagrams. Plots can be used for tests that can have a range of values so again would not be applicable to search filter performance. A flow diagram, however, could be considered as a method for presenting search filter performance.

Comparing the results of search filters

Systematic reviews of the DTA literature are complex, largely because of the variability (heterogeneity) between studies in terms of the reference standards that have been used and the populations that have been tested. When heterogeneity exists, meta-analysis is not recommended and a narrative approach is advised using graphical presentations such as forest plots and ROC space plots.

In the search filter literature, a variety of approaches have been adopted to test search filters using different search interfaces and so heterogeneity is likely to be present between filters. There have been few systematic reviews undertaken in the search filter literature and these have tended to adopt a different approach from that taken in the DTA literature. Although DTA reviews frequently compare studies that have evaluated the performance of one index test against the performance of the same reference standard but in different samples, search filter reviews published to date compare several search filters using both the same reference standard and sample (review B). In this situation, synthesising the results is not applicable; rather, we can directly compare performance between filters. These reviews have tended to display the results only in tabular form but ROC space plots or paired forest plots would be highly appropriate for displaying these comparisons. Displaying the results using graphs may convey them more effectively and assist users to choose between filters.

Conclusions

Guidance on conducting and analysing the results of DTA studies is applicable to several aspects of search filter research. The identification of a representative sample of records, of sufficient size and using a standardised approach will assist in producing robust and generalisable results. Although appropriate performance measurements are generally reported, the greater use of some graphical presentations may facilitate the dissemination and interpretation of results.

How do searchers choose search filters? (review D)

Objectives

The objective of this review was to identify any published research into how searchers (information specialists, librarians, researchers and clinicians) choose search filters based on the information presented to them.

Methods

Studies were eligible for inclusion if they reported criteria or methods that searchers used to choose filters, for example:

- the characteristics of the filter, such as how the filter was designed, what performance measurements were used and the currency of the filter
- how searchers appraised the filter designs, for example, did they use the ISSG critical appraisal tool,[4] the Canadian Agency for Drugs and Technologies in Health (CADTH) tool[3] or other methods to appraise search filters to inform their choice

- whether or not searchers asked for advice from others on the choice of filters, including colleagues, recognised experts in the field (such as members of the ISSG or the McMaster Hedges project team) or other professional networks
- where searchers found the filter; for example, did they choose the filter because they found it in a source they regarded as 'reputable' (such as MEDLINE/PubMed or the ISSG Search Filters Resource) or in published guidance documents [such as those produced by the Centre for Reviews and Dissemination (CRD)[69] or Cochrane[114]].

Studies were excluded if they were not specifically about search filter choice or were in languages other than English. Studies from any discipline were eligible.

Although there is a large volume of literature on resource selection, this is not directly applicable to this very specific type of tool selection. At the protocol stage we decided against searching for generic literature about resource selection 'choices' as this was likely to retrieve a large number of records with little or no direct relevance to the review question.

To identify relevant studies we searched databases in a number of disciplines including information science and health care. *Table 17* summarises the database and other resources searched to identify relevant studies.

The search strategy consisted of subject indexing (e.g. MeSH, Emtree) and free-text terms (in the title and abstract). It included search terms for 'searchers/information specialists' in combination with terms for 'choice/decision' and terms for 'methodological search filters'. No date or language limits were applied to

TABLE 17 Review D: databases and other resources searched

Resource	Interface/URL
MEDLINE (and MEDLINE In-Process & Other Non-Indexed Citations)	OvidSP
EMBASE	OvidSP
PsycINFO	OvidSP
Library, Information Science and Technology Abstracts (LISTA)	EBSCO*host*
Cochrane Methodology Register	The Cochrane Library/Wiley Online Library
SCI	ISI Web of Science
SSCI	ISI Web of Science
CPCI-S	ISI Web of Science
CPCI-SSH	ISI Web of Science
HTAi Vortal	http://vortal.htai.org/ (accessed 29 October 2010)
EUnetHTA	https://eunethta.fedimbo.belgium.be/ (accessed 1 November 2010)
HTA organisation websites: INAHTA, AHRQ, CADTH, CRD, CEDIT, AETS, DAHTA, IQWiG, OSTEBA, SBU[a]	Various (accessed 1–3 November 2010)
UK Health Libraries Group	www.cilip.org.uk/about/special-interest-groups/health-libraries-group (accessed 1 November 2010)
EAHIL	http://eahil.eu/ (accessed 1 November 2010)
US Medical Library Association	www.mlanet.org/ (accessed 1 November 2010)

AETS, Agencia de Evaluación de Tecnologías Sanitarias; CEDIT, Comité d'Evaluation et de Diffusion des Innovations Technologiques; CPCI-S, Conference Proceedings Citation Index – Science; CPCI-SSH, Conference Proceedings Citation Index – Social Science and Humanities; DAHTA, German Agency for Health Technology Assessment; EAHIL, European Association for Health Information and Libraries; INAHTA, International Network of Agencies for Health Technology Assessment; IQWiG, Institute for Quality and Efficiency in Health Care; OSTEBA, Basque Office for Health Technology Assessment; SBU, Swedish Council on Health Technology Assessment; SCI, Science Citation Index; SSCI, Social Science Citation Index.
a See *Appendix 4*.

the search. Full search strategies are listed in *Appendix 4*. Records were downloaded from databases and then imported into EndNote X5 bibliographic software (Thomson Reuters, CA, USA), which allowed categorisation and coding, as well as streamlining of the production of draft and final reports. Duplicate records were then removed.

The titles and abstracts of the records identified in the searches were assessed for relevance. The intention was to select those studies reporting how searchers make choices about search filters. Studies not specifically about search filter choice and studies in languages other than English were excluded.

We acknowledge that there has been a regrettable delay between carrying out the project, including the searches, and the publication of this report, because of serious illness of the principal investigator. The searches were carried out in 2010/11.

Results

In total, 2266 records were identified by the searches. *Table 18* shows the numbers of records by resource identified from the searches.

After the removal of duplicates, 837 records remained for assessment. The titles and abstracts of these 837 records were assessed for relevance and no records met the inclusion criteria (*Figure 9*).

Discussion

The search strategy used search terms relevant to systematic review methods ('search strategy', 'search filter', 'information specialist', 'choice/decision') and as a result a high proportion of the records identified

TABLE 18 Review D: numbers of records identified from various resources

Resource	Number of records identified
MEDLINE (and MEDLINE In-Process & Other Non-Indexed Citations)	638 (14)
EMBASE	824
PsycINFO	30
Library, Information Science and Technology Abstracts	164
Cochrane Methodology Register	57
SCI	420
SSCI	100
CPCI-S	14
CPCI-SSH	5
HTAi Vortal	0
EUnetHTA	0
HTA organisation websites: INAHTA, AHRQ, CADTH, CRD, CEDIT, AETS, DAHTA, IQWiG, Osteba, SBU[a]	0
UK Health Libraries Group	0
EAHIL	0
US Medical Library Association	0

AETS, Agencia de Evaluación de Tecnologías Sanitarias; CEDIT, Comité d'Evaluation et de Diffusion des Innovations Technologiques; CPCI-S, Conference Proceedings Citation Index – Science; CPCI-SSH, Conference Proceedings Citation Index – Social Science and Humanities; DAHTA, German Agency for Health Technology Assessment; EAHIL, European Association for Health Information and Libraries; INAHTA, International Network of Agencies for Health Technology Assessment; IQWiG, Institute for Quality and Efficiency in Health Care; Osteba, Basque Office for Health Technology Assessment; SBU, Swedish Council on Health Technology Assessment; SCI, Science Citation Index; SSCI, Social Science Citation Index.
a See *Appendix 4*.

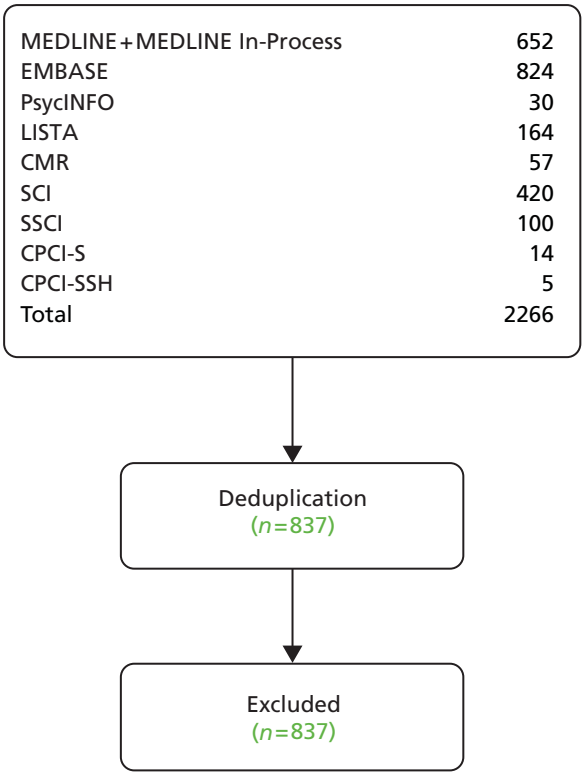

FIGURE 9 Review D: numbers of records retrieved and assessed for relevance. CMR, Cochrane Methodology Register; CPCI-S, Conference Proceedings Citation Index – Science; CPCI-SSH, Conference Proceedings Citation Index – Social Science and Humanities; LISTA, Library, Information Science and Technology Abstracts; MEDLINE In-Process, MEDLINE In-Process & Other Non-Indexed Citations; SCI, Science Citation Index; SSCI, Social Science Citation Index.

were systematic reviews, which typically report search strategies in their abstracts. In total, 48% (402/837) of the records assessed were Cochrane reviews, which report their methods in detail and whose abstracts tend to include search terms similar to those used in this search strategy. Many other non-Cochrane reviews were also identified for the same reason. This also explains the high number of duplicate records retrieved as Cochrane reviews were identified across most of the databases searched.

Studies about the creation, testing, evaluation and awareness of search filters were also identified because of the similarity of the search terms used in the strategy and those used in the bibliographic records. Other studies looked at search techniques for identifying study populations by age or sex; investigated the differences between databases and database interfaces; and discussed the growing importance of searching via the internet. In addition, a significant number of records were completely irrelevant, such as those about searching bioinformatics (genes, proteins) databases.

However, we did not identify any studies that had explored how searchers select search filters. The absence of studies was not unexpected, despite the fact that our searches were relatively sensitive and were undertaken across a wide range of resources (including databases covering health care and information science as well as HTA organisation websites).

It was decided when developing the protocol that, given the resources available for this project, it would not be possible to undertake broader searches to identify research about how searchers or information specialists (including librarians) make choices about the resources/tools they use. It was felt that this literature would be very large as it would include library stock selection, database selection and other situations in which informed choice is required. It may be that this literature could suggest how information seekers choose between tools. The literature would not be specific, however, to the choice of search filters and might be qualitatively different as many stock selection decisions may be governed by factors such as cost and subject coverage rather than sensitivity and precision.

There is literature about the development and quality of search filters, as well as research comparing published filters, but we did not identify any studies reporting the use and choice of filters by searchers in practice. A survey about the awareness of search filters among searchers was published in 2004 and, although awareness of filters was relatively high at that time, usage was still low.[5] Since that questionnaire was undertaken, the promotion of search filters through the ISSG Search Filters Resource, through training courses conducted in the UK, the USA and elsewhere and through the increasing numbers of published filters may have increased awareness and usage by searchers. We have not identified any current published evidence, however, to support this. Investigations of how searchers are choosing filters seem not to have been published.

How do clinicians choose between diagnostic tests? (review E)

Introduction

Database searchers have access to a range of methodological search filters that have been designed to retrieve records relating to studies that employ a particular research design. It is unclear, however, what factors influence the choice of an appropriate filter. As search filters can be viewed as analogous to diagnostic tests (as outlined above), it is hypothesised that the factors that lead clinicians to choose between diagnostic tests or health-care organisations to choose between screening tests might offer insights into how searchers do, or might in the future be encouraged to, make choices about search filters.

Objective

To identify and summarise evidence, in a narrative review, on factors that influence clinicians' choice between diagnostic tests.

Methods

Evidence for this review was obtained from literature searches of the major health-care databases and consultation of national screening programme websites. MEDLINE, MEDLINE In-Process & Other Non-Indexed Citations and EMBASE were searched in March 2011 and CINAHL, PsycINFO and Applied Social Sciences Index and Abstracts (ASSIA) were searched in June 2011. The search strategies that were used are reproduced in *Appendix 5*. No date restrictions were applied but a pragmatic decision was taken to search only for English-language publications. Reference lists of relevant studies were scrutinised and citation searching of key articles was undertaken in Scopus and ISI Web of Knowledge. Results were downloaded into Reference Manager 12 (Thomson ResearchSoft, San Francisco, CA, USA). Titles and abstracts were screened and full-text copies of all studies deemed to be potentially relevant were obtained and assessed for inclusion by one researcher.

We acknowledge that there has been a regrettable delay between carrying out the project, including the searches, and the publication of this report, because of serious illness of the principal investigator. The searches were carried out in 2010/11.

Inclusion criteria

- Studies that report how clinicians choose between diagnostic tests and what factors influence their decisions.
- Screening programmes that provide criteria for the selection of screening tests.

Exclusion criteria

- Studies that report on any factors influencing test ordering decision behaviour without reference to test choice.
- Studies that consider the decision whether or not to order one particular test.
- Studies that report interventions designed to influence test ordering behaviour.
- Studies written in languages other than English.

Data extraction

For studies meeting our criteria, the following information was collected:

- research method(s) used to elicit data
- clinical discipline of participants and setting
- clinical condition or disease and diagnostic tests from among which clinicians made their choice
- factors implicated in clinicians' choice.

Results

The electronic searches retrieved 1559 records after deduplication (*Figure 10*). Titles and abstracts were screened and 47 records were selected for full-text assessment. Seven studies met the inclusion criteria.[115–121] *Table 19* provides details of the included studies. The references and citations of these seven publications generated an additional 38 articles for further checking, none of which met the inclusion criteria.

Studies were excluded for a variety of reasons. One-quarter (10/40) of the excluded studies considered the reasoning that underpins diagnostic decisions, mainly factors that can lead to errors and suboptimal diagnostic strategies, and one-quarter (10/40) surveyed the use of a range of tests for different conditions. Six articles examined factors that influence the diagnostic process or adopted strategy, characterised by a stepwise series of hypothesis testing using information from a variety of sources and series of tests. These included symptoms elicited from patients, patient and physician characteristics and structural issues.

Other reasons for exclusions were examination of patient choice or compliance ($n = 4$), use of interventions designed to influence test ordering behaviour ($n = 2$) and use of an economic model to assess screening strategies ($n = 1$). An additional two articles did examine test choice but did not elicit the reasons involved. *Appendix 6* provides details of the excluded studies together with the primary reason for exclusion.

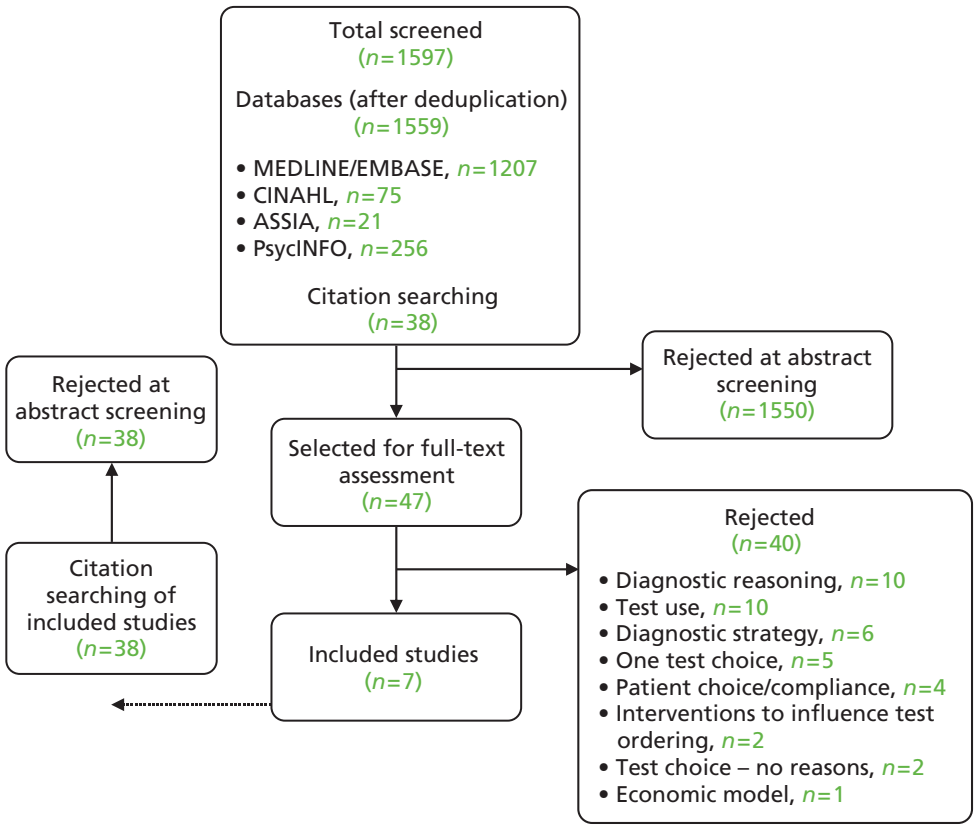

FIGURE 10 Review E: numbers of records retrieved and assessed for relevance.

TABLE 19 Review E: included studies

Study	Subjects; study location	Method	Tests	Results
Jha 2010[115]	Radiologists ($n = 62$), emergency physicians ($n = 52$); USA	Online questionnaire asking which diagnostic tests from a list participants would use to detect pulmonary embolism and why	CT scan, V/Q scan, angiogram, Doppler ultrasound, MRI, chest radiography	96% of emergency physicians and 90% of radiologists chose CT as the first-line investigation. Participants cited accuracy (90% and 95%), access (85% and 71%) and 24-hour interpretation (69% and 45%) as the main reasons for choosing this test. Non-availability of the other tests was not considered important
McGinnis 2009[116]	Physiotherapists ($n = 11$); USA	Qualitative – grounded theory approach. Participants were interviewed and undertook sorting activities of different assessment approaches	Balance assessment tests	Experience was the primary influence on choice. Patient level of function also contributed. Few valued the psychometric properties of the tests. The perceived value of information gathered mattered more than testing time. Tests with numerical scores were chosen for documentation purposes
Perneger 2002[117]	1994 physicians, 59% response ($n = 1184$); Switzerland	Mailed survey. Physicians were presented with a choice between two tests: test A, to be given to the whole population (1000 lives saved) vs. test B, which was a better, more expensive test to be given to half of the population (1100 lives saved)	Cancer screening tests (hypothetical)	75% opted for test A. Test B would be more acceptable if a clinical decision was involved in who received it
Sox 2006[118]	1502 paediatricians randomly selected, 49.7% response rate ($n = 653$); USA	Participants were mailed a questionnaire containing one of two clinical vignettes and were asked to choose between several tests for the vignette. Subjects were randomised to receive no further information (control), DTA performance (TC), DTA performance with a non-technical explanation (TC defined)	Culture, DFA test, PCR test	Significantly more participants in the TC and TC defined groups chose PCR (best performing test) than participants in the control group (73% vs. 71% vs. 21%) but this did not affect clinical management
Stein 2011[119]	Consensus group of experts in the field of pulmonary embolism ($n = 33$); multinational	Survey on the diagnostic management of pulmonary embolism	CT venography, CT angiography, SPECT, V/Q scan, ultrasound	Factors influencing opinions included test performance (sensitivity, specificity), risk of adverse events such as radiation exposure, added benefit set against resource use, patient factors (age, sex) and chest radiography results

TABLE 19 Review E: included studies (*continued*)

Study	Subjects; study location	Method	Tests	Results
Wackerbarth 2007[120]	Primary care internists and family physicians ($n = 66$); USA	Participants underwent semistructured interviews. Transcripts were reviewed and decision heuristics were developed: when to recommend screening; what type of screening	FOBT, flexible sigmoidoscopy, colonoscopy, double-contrast barium enema	Choice of screening test was influenced by patient characteristics (age, family history), health insurance coverage, patient acceptance and presenting symptoms
Zettler 2010[121]	894 primary care physicians randomly selected, 52% response rate ($n = 465$); Canada	Participants were mailed a survey asking which screening test they would use	FOBT, colonoscopy, flexible sigmoidoscopy, double-contrast barium enema	Significant association between screening choice and perceived test sensitivity, perceived cost-effectiveness and mortality reduction but not waiting times

CT, computed tomography; DFA, direct fluorescent antibody; FOBT, faecal occult blood test; MRI, magnetic resonance imaging; PCR, polymerase chain reaction; SPECT, single photon emission computed tomography; TC, test characteristics; V/Q, ventillation/perfusion.

Of the seven studies that met the inclusion criteria, none was set in the UK. Four studies were set in the USA,[115,116,118,120] one was set in Canada,[121] one was set in Switzerland[117] and one was multinational.[119] Information from the clinicians was obtained by survey ($n = 3$[117,119,121]), questionnaire ($n = 2$[115,118]) or interview ($n = 2$[116,120]) and the number of participants ranged from 11[116] to 1184.[117] Three studies looked at cancer screening tests (two for colorectal cancer),[117,120,121] two at imaging tests for pulmonary embolism,[115,119] one at balance assessment tests[116] and one at tests to diagnose pertussis.[118]

Four studies mentioned high test performance as a reason in support of clinician choice. In the study by Jha *et al.*,[115] 90% of emergency physicians and 95% of radiologists who responded to a questionnaire cited test accuracy as a reason for test choice. Both Stein *et al.*[119] and Zettler *et al.*[121] noted that perceived test performance was a factor in decision-making whereas Sox *et al.*[118] reported that 70% of participants who had received information on DTA performance chose the best-performing test compared with 21% of controls who had not received this information. One further study, which interviewed physiotherapists about balance assessment tests, found that the perceived value of information gathered was a deciding factor in clinician choice of test rather than the psychometric properties of the assessment tests.[116]

Two studies reported economic factors: the perceived cost-effectiveness of colorectal cancer screening tests[121] and the perceived added benefit as set against resource use of various diagnostic tests for pulmonary embolism.[119] One further study looked at the influence of equity in physician choice.[117] The participants were asked to choose between one test given to the whole population and a better (in terms of lives saved) and more expensive test given to half of the population. Three-quarters (75%) opted for the universal test although the better, more expensive test was seen as being more acceptable if clinical factors determined who would receive it.

Two studies reported patient characteristics as factors influencing test choice. Stein *et al.*[119] mentioned age and sex whereas Wackerbarth *et al.*[120] identified family history as an influencing factor for screening at an earlier age. Patient acceptance of the proposed tests and whether or not the tests were covered by patients' insurance coverage were also mentioned.[120]

Other factors considered were clinician experience (McGinnis *et al.*[116] reported this as the primary influence on test choice for balance assessment), mortality reduction[121] and adverse events, primarily in terms of

radiation exposure.[119] The study by Jha *et al.*,[115] which took place in an emergency department, found that ready access to the test and whether or not 24-hour interpretation support was available were the two most frequently reported factors after test performance.

In addition to the studies identified in the review, information on selection criteria for four screening programmes was identified (*Table 20*). Three of the four screening programmes that provided information were national, set in the UK,[122] USA[123] and Australia.[124] The fourth, providing criteria for cancer screening, was produced by the World Health Organization.[125] Most programmes identified high test performance in terms of sensitivity,[123–125] specificity,[124,125] PPV[124,125] and/or NPV[124,125] as important. The UK programme[122] stipulates that the test should be precise and that the distribution of test values in the target population should be known and a suitable cut-off level should be defined.

Other characteristics listed included being safe,[122,124,125] being reliable,[124] having been validated,[122,124] easy to administer[122,124] and being acceptable to the target population.[122,124,125] All of the programmes consider factors other than test performance. The effectiveness of undertaking a screening programme, in terms of morbidity and mortality reduction, should be established,[122,123,125] with effective identification of disease at an early disease stage[124] and the availability of effective treatment.[123,125] The condition under investigation should be sufficiently prevalent[123,125] so that a screening programme can be effective. The UK programme[122] adds that an agreed policy of further diagnostic investigation and disease management should have been agreed. Both the UK[122] and the USA[123] programmes mention that the perceived benefits of the screening programme should outweigh any harms resulting from screening and treatment.

TABLE 20 Review E: reports from national screening programmes

Report	Details
UK National Screening Committee 2011[122] (criteria for appraising the viability, effectiveness and appropriateness of a screening programme)	Criteria to be met: a simple, safe, precise and validated screening test; the distribution of test values in the target population should be known and a suitable cut-off level should be defined and agreed; the test should be acceptable to the population; an agreed policy on the further diagnostic investigation of individuals with a positive test result and on the choices available to those individuals; evidence from high-quality RCTs that the screening programme is effective in reducing mortality or morbidity; benefits from the screening programme should outweigh the physical and psychological harms (caused by the test, diagnostic procedures and treatment)
US Preventive Services Task Force 2008[123] (procedure manual)	Criteria to be met: assess net benefit; prevalence of the condition; sensitivity of the test; effectiveness of early treatment; reduction in morbidity/mortality; harms of screening; harms of treatment
Australian Population Health Development Principal Committee Screening Subcommittee 2008[124] (population-based screening framework)	Criteria to be met: effective at detecting early-stage disease, valid, safe, reliable, high sensitivity, high specificity, high PPV, high NPV, easy to perform and interpret, acceptable to the target population
World Health Organization 2011[125] (screening for various cancers)	Fundamental principles: the target disease should be a common form of cancer, with high associated morbidity or mortality; effective treatment, capable of reducing morbidity and mortality, should be available; test procedures should be acceptable, safe and relatively inexpensive; the following factors should be taken into account: sensitivity – the effectiveness of a test in detecting a cancer in those who have the disease, specificity – the extent to which a test gives negative results in those who are free of the disease, PPV – the extent to which subjects have the disease in those who give a positive test result, NPV – the extent to which subjects are free of the disease in those who give a negative test result, acceptability – the extent to which those for whom the test is designed agree to be tested

Discussion

From this overview it seems that there is limited evidence to clarify how clinicians choose between diagnostic tests. What evidence there is suggests that test performance is the main factor that informs their choice. It has been reported, however, that a substantial proportion of clinicians have an inaccurate understanding of test performance parameters and apply them inaccurately[126–131] and so it may be the case that choices are being based on false assumptions. Other factors mentioned in more than one study were the pretest probability of having the condition, as defined by patient characteristics, patient acceptance of the test and the costs involved in carrying out the test, which are factors that are not readily transferable to the search process. Additional attributes reported related to the particular scenario being investigated: the harmful effect of radiation when imaging tests were being considered and the need for immediate testing and interpretation in an emergency department were important criteria in two studies.

The screening programmes also valued high test performance but add that a test should have been proven to be valid and reliable. Furthermore, the screening committees set other criteria to ensure the effectiveness of public health programmes: the prevalence of the target disease or condition as well as whether or not there is effective disease management and treatment available. In a screening setting, where patients are asymptomatic, acceptability was mentioned as crucial by three of the screening programmes and the need to evaluate benefits against harms was also considered to be an important criterion.

Conclusion

From the very limited evidence available in a clinical setting, it is difficult to gain much insight into how searchers might make choices about search filters. Diagnostic test performance (perceived or known) was the most frequent factor mentioned and is the main factor that is readily applicable to search filter choice. However, it may be beneficial to provide additional explanatory information when reporting search filter performance to ensure that searchers make choices based on an accurate understanding of test performance parameters.

Chapter 3 Interviews

Aims

Interviews were carried out to inform the development of the questionnaire and the subsequent pilot website and guidance sections of this report. The aim of the interviews was to learn how search filters are used by information professionals working in NICE and organisations affiliated to NICE.

Methods

A semistructured interview protocol was developed. Information professionals working for NICE, NICE Collaborating Centres and NICE Evidence Review Groups (ERGs) were contacted and asked if they would be willing to be interviewed.

A total of 12 interviews were carried out, capturing the views of 16 information specialists drawn from 14 organisations within the NICE family (NICE, four NICE Collaborating Centres and nine NICE ERGs) (*Table 21*).

None of the senior NICE information staff interviewed had roles that involved operational information retrieval work. The current roles of NICE staff focused on providing quality assurance and guidance for their teams. All of the NICE staff interviewed had considerable searching experience from previous roles.

The interviews lasted for approximately 45 minutes and all but one were conducted by telephone; the interview not conducted by telephone was conducted face to face. The interviews took place between 1 January 2009 and 3 March 2009.

Findings

Databases used by interviewees

The interviewees use or have used a range of databases, many of which are health related (*Table 22*). Other databases mentioned were project specific and included databases that focused on social care, transport, criminology and humanitarian aid.

Interviewees' use of search filters

Circumstances under which NICE searchers did not tend to use search filters included the following:

> *For short clinical guidelines, the team only use search filters on the rare occasions when the PICO [population, intervention, comparison and outcome] is restricted to study design.*
>
> *Filters do not work very well when searching for diagnostic studies.*

TABLE 21 Numbers of interviews and interviewees

Number	
Interviewees per interview	Interviews conducted
1	10
2	1
4	1

TABLE 22 Health databases used by the interviewees

Database	Interviewees from	
	NICE	NICE Collaborating Centres and ERGs
MEDLINE	4	8
MEDLINE In-Process & Other Non-Indexed Citations	1	1
MEDLINE Daily Update	1	
EMBASE	4	8
EMBASE Alert		1
The Cochrane Library databases	3	6
CDSR	1	
CENTRAL	1	
DARE	1	
HTA database	1	
NHS EED	1	
AMED	1	
CINAHL	4	3
Clinical trials databases and trials registers	1	1
Guidelines resources		1
HEED	1	1
HMIC		1
PsycINFO	1	6
Scopus		1
Social Policy in Practice	1	1
Transport	1	
Web of Science		1

AMED, Allied and Complementary Medicine Database; HMIC, Health Management Information Consortium.

There is only a small volume of literature relating to new procedures/interventions, so filters are not necessary.

Searches carried out at the point in time when products get a CE [Conformité Européene] mark (or before), tend to be internet-based as any publications are very new and may not yet be included in databases.

The ERGs' use of search filters for NICE work was limited because:

Single Technology Appraisals involved a review of the work of organisations submitting to NICE. The ERG staff only developed searches to test the searches carried out by the submitting body.

Multiple Technology Appraisals (MTAs) are very PICO [population, intervention, comparison and outcome]-driven.

However, the occasions when filters are used by an ERG included:

> *When carrying out searches for systematic reviews that include RCTs.*
>
> *To help focus the question further than PICO [population, intervention, comparison and outcome] permits, to make the project manageable in terms of record numbers retrieved.*
>
> *With projects looking at a single study type which are usually small projects with limited resources.*
>
> *To build searches to answer guideline questions, except on the occasions where search results were small in number.*
>
> *To identify economic evidence.*
>
> *To carry out limited focused searches.*

The filters that interviewees said that they used were:

> *Cochrane RCT filters and RCT filters [unspecified].*
>
> *Diagnostic test accuracy filters.*
>
> *Qualitative filters [drafted by the interviewee].*
>
> *Filters produced by HIRU [Health Information Research Unit]/the McMaster Hedges Team.*

Where would you look for a search filter?

Interviewees provided several responses to this question:

> *CRD website/blog/ISSG search filters page/InterTASC website [note that the last two sites mentioned here are the same].*
>
> *Would post a question to discussion lists.*
>
> *Look in the Cochrane Handbook.*
>
> *Speak to colleagues.*
>
> *Consult an in-house methodology database.*
>
> *Consult an in-house search manual.*
>
> *Look at methods used in previous project.*

Developing and amending search filters

Some interviewees were comfortable with translating filters for use in different databases and some were not. Interviewees were comfortable translating MEDLINE search filters for use in other databases but were not comfortable translating non-MEDLINE filters for use in other databases. In the absence of objectively derived filters, however, interviewees said that they would have translated non-MEDLINE filters to run them in other databases.

Some interviewees said that to identify qualitative research they would tend to write/amend their own filters. Some respondents said that they would amend filters for scoping searches. A number of

respondents noted that filters need to be written (or adapted) on a review-by-review basis depending on what was needed from the search. Several respondents indicated that they would adapt filters occasionally, for example if a filter was too sensitive they would take out a few lines to make it more specific.

Reporting the use of search filters

A number of approaches were reported around the documentation of the use of search filters:

Citing the search filters used and reporting if amendments had been made to an existing filter or if the strategy had been based on an existing filter.

Writing search strategies up fully without explicitly citing the filters used.

Including search filters as part of published strategies but not explicitly identifying them.

Documenting the use of all agreed filters, amendments and the rationale behind the amendments.

Keeping a record of search strategies but not describing them when the strategy is not written up for publication.

Using a process document which included a section about which filters have been used; document the filter when there is a need to justify the use of a filter.

Methods of keeping up to date

Interviewees' attitudes towards keeping up to date ranged from 'Difficult – something that is always on the "to do" list' to 'As we are such a small community I feel that it is unlikely that important information about a new filter will be missed'. Interviewees reported using the following methods to keep up to date:

A NICE internal current awareness bulletin.

E-mail lists and specialist groups (e.g. Cochrane IRMG [Information Retrieval Methods Group], HTAi IRG [Information Resources Group], National Library of Medicine list for MeSH changes and updates).

Meetings {e.g. ISSG, groups in the wider NICE family (e.g. NCC [National Collaborating Centre] Information Specialist Network meetings)}.

Websites (e.g. ISSG and McMaster Hedges team).

Conferences (e.g. Cochrane, HTAi).

Journal publications.

Setting up a citation search.

Choosing between filters

Interviewees reported a wide variety of actions that they might take when choosing between two or more filters, including:

In cases where two search filters appear similar (in terms of sensitivity and specificity) I tend to take the good parts from each and test to see that the results still include benchmark papers and then make sure client is happy with the approach.

Sensitivity and specificity figures can give a guide but they are still reporting the results from that instance. They may have been combined with a specific topic or used in a specific context and will still have to be investigated for appropriateness.

I would test for sensitivity and specificity. The final choice, however, is still arbitrary, relying on gut-feeling and the requirements of the specific project.

Test against a set of target references.

If I had sufficient time would try both and test results against each other; if there was a lack of time I would use the most current.

I assess the methods used to develop the filter and the extent to which it matched needs of the search (sensitivity, precision, a mixture of the two).

I would back up my decision with academic literature.

I run both filters and compare the results to see where there are gaps/duplications in retrieval between the two and to see which retrieves the more relevant papers.

I try both – I use my gut feeling rather than anything formal.

Provenance – I judge according to who developed the filter.

Look on ISSG website to see if anyone has completed an appraisal.

Search testing is pragmatic/intuitive rather than being a formal scientific process (there is not enough time to do this).

I would like someone to be quite directive about which are the best filters to use in different situations and be able to quickly see how these filters have been evaluated and how decisions have been reached (e.g. as in the Cochrane Handbook).

As a junior information specialist, the decision on which filter to use is made by senior colleagues.

It would be easier to choose if the Collaborating Centres and NICE were using the same filters and then informed everyone when changes/updates have occurred.

The YHEC [York Health Economics Consortium] 'Getting the best out of search filters' training course has been useful information to help critically appraise filters.

What would help you choose between filters?

Interviewees provided a range of responses when asked what would help them choose between filters:

The interpretation of the filter in simple terms – such as power calculations, statistical methods etc. that are difficult to understand, particularly for those with limited time – a synopsis would be a great help.

It is difficult to fully understand all of the complicated technical methods used to devise and test search filters. There is an element of trusting the researchers involved – I can critically appraise to a certain level but not entirely.

A summary documenting sensitivity/specificity would help to choose between filters, although this might be subjective (e.g. a document would be good for one search but not for another).

Sensitivity and specificity are important.

Some measure of rating would be useful but I would need to have confidence in whoever had carried out the rating.

Benefits of filters

Interviewees said that the main benefits of filters were that they could target the results of searches and reduce the volume of literature retrieved. It was also mentioned that the use of established filters (e.g. the Cochrane RCT filters), which have been evaluated and tested, reflected well on search quality. Additionally, using filters means that the searchers can benefit from someone else's expertise and time spent developing the filters.

Limitations of filters

Interviewees expressed a range of concerns about search filters:

There is always a chance something has been missed.

Filters still identify a lot of irrelevant records.

Poor indexing doesn't help searching.

Few, if any, are used appropriately.

If there is a mistake, it will be replicated through all searches/databases.

Transferability can be a problem, e.g. the Fleming qualitative filter was originally devised for use in nursing topics and probably works fine there but it was not appropriate for a diagnostic type study.

Not many filters are reliable, there are only a few databases that you can use them in and databases keep changing so it is important to check that the filter is up-to-date.

A filter gets published, people talk about it and it gets known and it starts getting used. But negative results/experiences tend not to be talked about or published and therefore there can be bias.

Areas where filters are needed/existing filters need to be improved

Interviewees were asked if there were any topics for which filters are needed or any filters that could be improved. The responses included the following:

Population age.

HRQoL (including topic-specific instruments).

Tested filters for observational studies.

Epidemiology.

Diagnostics.

Adverse events and safety issues.

Prognostic filters.

Qualitative research filters.

An improvement to the diagnostic accuracy filter.

Other comments

Interviewees were asked for other final comments and responded with both general and specific points:

The methods behind derivation of filters are impenetrable, so there is a certain amount of trust involved in using them. But this is an improvement on pragmatically deriving a study design filter from scratch.

Databases need some/better coding for SRs [systematic reviews] and DTAs [DTA studies].

PRISMA [Preferred Reporting Items for Systematic Reviews and Meta-Analyses] guidelines about reporting search strategies need to be reviewed.

There is a need for academics to recognise the importance of searching in its own right – this might be helped if information specialists were to routinely write-up protocols and include academic arguments for the approach they took.

Perhaps developers of similar filters could work collaboratively/liaise with one another to see if there is really a need for two or more filters which (appear to) carry out the same role.

There are issues with different database interfaces. For example, a filter devised for use in Ovid is likely to work very differently if translated into another interface, such as EBSCO (or Web of Science or Dialog DataStar, etc.).

More education on what filters can and can't do is important as there are still examples of filters being used incorrectly, for example, an economic filter being used in NHS EED, an RCT filter used in CENTRAL.

Filters are needed for more databases, rather than more filters for MEDLINE (and EMBASE).

There are problems with EMBASE and the number of Emtree terms attached to the records leading to the retrieval of more irrelevant records.

It would be useful if the ISSG Search Filter Resource website indicated when something new had been added.

Patient experience/issues filter (SIGN) [Scottish Intercollegiate Guidelines Network] needs to be disaggregated – it is over 200 lines long.

Discussion

The interviewees were information specialists involved in searching as part of NICE or an organisation providing support for the development of NICE guidelines and technology appraisals. It seems likely that the majority of the interviewees were experienced searchers but fairly senior. This means that some of our interviewees were no longer searching currently on a daily basis and using filters but had done so in the past. Nevertheless, the views of senior staff are valuable as they represent the staff that are setting search standards and policies within NICE. It should be noted that some interviews were undertaken in groups and this could have influenced the responses.

The interviews revealed the wide range of searching tasks that are undertaken in the NICE context and the various points at which search filters can be used. However, there were many tasks for which search filters were not considered necessary or appropriate. The use of search filters seemed to be linked to reducing large numbers of records, introducing focus and assisting with searches that are focused on a single study type.

The Cochrane RCT filter was most often cited as a filter in common use as well as filters produced by the McMaster Hedges team. The methods used to identify filters were various but the most frequently mentioned resource was the ISSG Search Filters Resource.[6] This is likely to reflect the high profile given to this resource by the NICE family of information specialists.

Interviewees' practices when using, adapting and reporting search filters were far from uniform, possibly indicating an absence of accepted published formal guidance on these issues. In the absence of guidance, variations in practice can occur.

Current awareness methods were varied and extensive. Interviewees were stretched in terms of keeping informed about search filter developments because of time limitations. This is likely to be because this is only one of many aspects of the rapidly evolving field of information retrieval methods.

When choosing filters we observed that interviewees were trying to make judgements around the relative sensitivity, specificity and precision of the filters but were conscious of factors impeding this. These factors included time constraints and knowledge gaps. The reference by interviewees to 'gut feeling' shows the relatively informal and pragmatic nature of search strategy testing and the absence of formal assessment or comparison tools to remove the necessity of relying on 'gut feeling'. Some interviewees expressed a desire for more guidance on the best filters to use or chose filters based on the authorship of the filter. This willingness to rely on the judgements and recommendations of others possibly reflects both a lack of time and a perception of an absence of the required skills to make informed judgements. Some desire for standardisation or guidance within the NICE family was also expressed.

Interviewees expressed their opinions on how making decisions about search filters could be assisted. These opinions were focused on making information about filters less technical and more user-friendly and offering 'bottom lines' or ratings. A synopsis of the interpretation of a filter was an additional feature that was suggested.

The disparity between the respondents' perceptions of the benefits and their perceptions of the limitations of search filters was marked. The respondents identified far more limitations than benefits and the limitations (poor precision, indexing weaknesses, filters created for a few key databases only) reflected the complex nature of searching, which filters alone cannot be expected to resolve.

However, the promotion of search filters as a tool could be improved by providing more guidance on best practice, summarising filters in non-technical, user-friendly ways and providing training on search- and filter-related issues. There appears to be demand for the development of filters for a range of methods areas and for a range of databases, but the limiting factor seems to be a lack of resources to develop such filters.

Chapter 4 Questionnaire

Sections of this chapter have been previously published in *Health Information and Libraries Journal*. Reproduced with permission from Beale *et al.*[132] © 2014 The authors. Health Information and Libraries Journal © 2014 Health Libraries Journal. Health Information & Libraries Journal, 31, pp. 133–147.

Questionnaire methods

A questionnaire survey was developed to obtain information on searchers' knowledge of and use of search filters. The questions were based on findings from the reviews and the interviews that had already been undertaken as part of this project. The questionnaire (see *Appendix 1*) was made available on the York Health Economics Consortium (YHEC) website.

Invitations to participate in the questionnaire survey were sent to seven e-mail lists:

1. LIS-MEDICAL (1523 individuals belonging to an open discussion list for members of the UK medical and health-care library community and other interested information workers)
2. irmg@lists.cochrane.org [204 subscribers belonging to the open discussion list of the Cochrane IRMG (Information Retrieval Methods Group)]
3. issg@lists.shef.ac.uk (subscribers are information specialists who work for the ERGs providing services to NICE and other associated individuals)
4. isg-informationresources@htai.org (subscribers are information specialists who are members of the HTAi organisation and other associated individuals)
5. Campbell IRMG (30 subscribers belonging to the Campbell IRMG and other associated individuals)
6. Cochrane Trials Search Co-ordinators (TSCs) (100 members of the Cochrane TSCs e-mail discussion list – now known as Cochrane Information Specialists)
7. EAHIL-L@MAILTALK.AC.UK [1000 members of the discussion list of the European Association for Health Information and Libraries (EAHIL)].

The invitation e-mail provided some background to the project and a link to the electronic questionnaire. To assist with completion, the e-mail also contained details of how to obtain a Microsoft Word 2010 (Microsoft Corporation, Redmond, WA, USA) version of the questionnaire for those who did not wish to complete the questionnaire online.

Additionally, short notifications were posted on Twitter (www.twitter.com; Twitter, Inc., San Francisco, CA, USA) and on the YHEC Facebook page (www.facebook.com; Facebook, Inc., Menlo Park, CA, USA), asking interested individuals to contact the YHEC for a link to the questionnaire survey.

The survey was available for completion during a 4-week period (22 July–18 August 2011), with e-mail reminders sent out 1 week before the final deadline.

In total, 90 survey responses were returned. It was not possible to calculate a response rate as it was not known how many individuals were members of more than one list, nor was it possible to determine the number of individuals who were alerted to the survey via the Twitter or Facebook messages.

Questionnaire results

What is your job title?

Forty-three different job titles were provided. Seventy of the 88 respondents who answered this question (79.5%) reported a job title that included the word 'library', 'librarian' or 'information'.

The remaining respondents reported one of the following job titles (two respondents did not answer this question):

- Assistant Professor
- Associate Scientist
- Consultant Physician and PhD candidate
- Director, systematic review research unit
- e-resources Co-ordinator
- Health Communication Specialist
- Learning Resources Officer
- Medical Documentalist
- Research Assistant
- Research Fellow
- Senior Lecturer
- TSC.

Over 75% of the respondents worked directly in information or library services, with the remaining respondents holding positions in which research and information finding would seem to be a key aspect of the role and knowledge of search filters could be assumed.

How long have you been searching databases such as MEDLINE?

The questionnaire was completed by experienced searchers, all with a minimum of 1 year's experience of database searching and with nearly half (48.9%; 44/90) having > 10 years of database searching experience (*Table 23*).

How often do you develop new search strategies as part of your work?

Three-quarters of questionnaire respondents (75.6%; 68/90) reported that they developed searches at least once a week and half of these said that they developed searches daily (*Table 24*).

For what purposes do you carry out searches within your organisation?

The questionnaire sought information on what types of searches were carried out. Respondents were presented with the following three options and were asked to tick all that applied:

- rapid searches to answer brief questions (78.9%; 71/90)
- scoping searches to estimate the size of the literature on a topic (81.1%; 73/90)
- extensive searches to inform evidence synthesis such as guidelines, systematic reviews and technology assessments (94.4%; 85/90).

TABLE 23 Length of time that respondents had been searching databases

Years of searching experience	Number of respondents	Percentage of respondents
< 1	0	0.0
1–5	15	16.7
6–10	29	32.2
11–15	15	16.7
16–20	15	16.7
≥ 21	14	15.6
No response	2	2.2
Total	90	100.0

TABLE 24 Frequency of developing new search strategies

Frequency of developing new search strategies	Number of respondents	Percentage of respondents
Daily	35	38.9
Once a week	33	36.7
Once a month	17	18.9
Less than once a month	5	5.6
Total	90	100.0

The most common searches that were carried out by respondents to the survey appear to be extensive searches to inform reviews and guidelines, but almost 80% of respondents reported that they also carried out rapid searches to answer brief questions and/or scoping searches.

Respondents also reported that they carried out searches for purposes other than those mentioned above. These were focused around teaching/education or were carried out in response to direct questions (*Table 25*).

Which databases do you search regularly?

Respondents were presented with a list of six databases (*Table 26*) (which are often cited for searches in HTA and systematic reviews) and were asked which they searched regularly. They were also asked to indicate any other databases that they use on a regular basis.

All respondents reported that they use MEDLINE and most (93.3%; 84/90) used The Cochrane Library databases. Over 75% of respondents indicated that they used EMBASE (77.8%; 70/90), nearly 75% used CINAHL (74.4%; 67/90) and > 60% (62.2%; 56/90) used PsycINFO. In total, 10% of respondents used HEED (10.0%; 9/90).

Other databases that were used by four or more respondents are documented in *Table 27*.

TABLE 25 'Other' searches reported by respondents

Other searches	Details
Teaching/education ($n = 3$)	'Demo search strategies' to assist students and academics to formulate strategies
	As part of teaching
	Searches for educational purposes (examples to use in teaching)
Responding to direct questions ($n = 10$)	General searches to answer queries more extensively than is the case for brief queries but less extensively than for systematic reviews
	Literature related to paediatrics
	Patient education queries, searches to support realist reviews, literature searches in support of medicolegal questions/lawsuits
	Analysis of a situation/bibliographic analysis/identifying trends
	Searches related to health research or policy-type questions
	Searches to help postgraduate students conduct literature reviews
	Competitive pipeline
	Searches to support literature reviews or clinical practice
	US FDA submissions
	Go/no-go feasibility studies for clinical trials

TABLE 26 Databases that are used regularly by respondents by frequency of citation

Database name	Number of respondents	Percentage of respondents
MEDLINE (including PubMed)	90	100.0
The Cochrane Library databases (CDSR, DARE, NHS EED, CENTRAL, HTA database)	84	93.3
EMBASE	70	77.8
CINAHL	67	74.4
PsycINFO	56	62.2
HEED	9	10.0

TABLE 27 Other databases searched by four or more respondents by frequency of citation

Database	Number of respondents who reported searching the database
Web of Science/Web of Knowledge	15
Scopus	8
Sociological Abstracts	8
Education Resources Information Center	6
ASSIA	5
Allied and Complementary Medicine Database	4
CRD	4
EconLit	4
Health Management Information Consortium	4
Turning Research into Practice	4

Have you ever used a methodological search filter?

Over 90% of respondents indicated that they had used methodological search filters (94.4%; 85/90); five respondents reported that they had not used a methodological search filter (5.6%; 5/90).

In what circumstances would you use methodological search filters?

Respondents were provided with five options to capture the circumstances in which they would use a methodological filter and were asked to tick all that they felt applied to their own situation (*Table 28*). Over 75% of respondents (76.7%; 69/90) indicated that they would use search filters for extensive searches carried out to find studies to inform guidelines or systematic reviews. Over 60% (61.1%; 55/90) indicated that they would use filters for rapid searches to answer brief questions and a similar number (58.9%; 53/90) said that they would use filters for scoping searches to estimate the size of the literature on a topic.

Respondents provided many other reasons for using search filters. Some related to developing the search strategy, some to the type of research that the search was informing and some to specific objectives:

- to practise search techniques
- to begin to identify MeSH and text words to use in developing a strategy
- if the customised limits provided by the databases cannot be relied on
- to reduce the results to a manageable size/narrow down results

TABLE 28 Circumstances in which search filters are used

Circumstances in which search filters are used	Number of respondents	Percentage of respondents
Extensive searches to inform guidelines or systematic reviews	69	76.7
Rapid searches to answer brief questions	55	61.1
Scoping searches to estimate the size of the literature on a topic	53	58.9
Other	12	13.3
None of the above	7	7.8

Reproduced with permission from Beale *et al.*[132] © 2014 The authors. Health Information and Libraries Journal © 2014 Health Libraries Journal. *Health Information & Libraries Journal*, **31**, pp. 133–147.

- to locate research conforming to appropriate methodology to inform systematic reviews or other research/clinical practice
- to meet client need/interest
- in analysis of a situation/bibliographic analysis/identifying trends
- in health research and policy, especially questions related to economics/cost-effectiveness
- to monitor trends
- for drug trials
- to keep updated regarding competitors' clinical trials.

Do you always use a filter when providing searches for similar types of projects?

Just over one-third of respondents (37.8%; 34/90) indicated that they would always use a filter when providing searches for similar types of projects. Just over half (56.7%; 51/90), however, would not and five respondents (5.6%; 5/90) did not respond to the question.

Four respondents indicated that they use filters only as a starting point when developing strategies and two respondents said that they rarely used filters, with one explaining that a filter would be used only when the topic had been well covered and a quick search was required. The circumstances in which respondents would not use a filter can be summarised as follows:

- when the volume of literature is manageable (21 respondents)
- client preference/specification (eight respondents)
- when looking for multiple study designs (six respondents)
- when looking for DTA studies (one respondent)
- on questions that encompass social issues (e.g. the social determinants of health, as much of the research is qualitative) (one respondent)
- when searching the literature for information neither directly for nor oriented towards clinical practice (e.g. physiology research) (one respondent)
- if it is important to be sure of finding all relevant references (one respondent)
- depending on the topic – it is not always appropriate and when undertaking scoping searches it is not always useful to narrow down these searches at an early stage (one respondent)
- when not sure that the filter is sufficiently sensitive (one respondent).

Typical practice when using search filters

Respondents were presented with different options describing how they might typically use search filters. The majority of respondents (81.1%; 73/90) indicated that they used different filters depending on whether their search needed to be sensitive or precise. However, 11% (10/90) of respondents reported using the same filter irrespective of the search focus (*Table 29*).

TABLE 29 Typical practice with respect to search filters

Statement of typical practice	Number of respondents	Percentage of respondents
I use different search filters depending on whether my search has to be sensitive or precise	73	81.1
I use the same search filter irrespective of the focus of the search	10	11.1
Other	7	7.8
Total	90	100.0

If you had to find a methodological search filter for a specific study design, where would you look?

Respondents reported a range of resources that they would use to identify search filters for specific study designs. Some respondents reported using more than one resource. When the respondents used varying designations for the same resource, these have been grouped together, for example responses such as 'Haynes', 'Hedges team', 'HIRU' and 'McMaster' have been grouped together as denoting the output of the McMaster Hedges team. Although respondents searched a range of resources, the most frequently searched resource for filters for a specific topic seemed to be the Cochrane filters to identify RCTs in MEDLINE[133] (36.7%; 33/90).

Across a range of topics, the most widely reported filters were those produced by the McMaster Hedges team, which are included in many interfaces to MEDLINE, as well as the filters reported on the ISSG Search Filters Resource. Search filters for RCTs and systematic reviews were more frequently reported than filters for other study designs.

In terms of search filters to find guidelines, respondents reported:

- using no filters (five respondents)
- developing their own filters (four respondents)
- using PubMed clinical queries systematic reviews or clinical queries filters (four respondents)
- using Health Evidence Bulletins Wales filters (two respondents)
- using McMaster Hedges filters (two respondents)
- searching using 'practice guideline.pt.' in MEDLINE (two respondents)
- using Scottish Intercollegiate Guidelines Network (SIGN) filters (two respondents)
- using various guideline producers' or guidelines.gov filters (two respondents)
- using Guidelines International Network filters (one respondent)
- using ISSG Search Filters Resource filters (one respondent)
- not needing filters to search for guidelines (one respondent).

In terms of search filters to find economic evaluations, four respondents indicated that they do not use filters and nine indicated that they have developed their own or adapted published filters. Other economics filters used by respondents were:

- SIGN filters (five respondents)
- CRD (NHS EED) filter (four respondents)
- MEDLINE/PubMed built-in queries (four respondents)
- McMaster Hedges filters (three respondents)
- specific databases rather than filters (two respondents)
- CADTH filters (one respondent)
- Comparative Effectiveness Research filters (one respondent)

- Guidelines International Network filters (one respondent)
- Health Services Research Queries (one respondent)
- ISSG Search Filters Resource (one respondent)
- McKinlay filters (one respondent)
- William Witteman's filter (from the Toronto HTA) (one respondent).

In response to a question about the types of filters (other than those for RCTs, systematic reviews, DTA studies, prognosis and aetiology) that they might use, five respondents indicated that they did not use a filter when looking for other types of studies and four respondents reported that they devised their own filters. Two respondents would search for filters on the ISSG Search Filters Resource. Other respondents suggested specific filters for a range of topics.

How do you decide which filter to use?

Respondents replied to this question by selecting one or more options from a list of options (*Table 30*). Respondents reported that they generally used the available filters that best suited their purpose (56.7%; 51/90) or the filters that were already available in the database being searched (53.3%; 48/90).

Respondents also noted other approaches that they used to help them decide which filters to use, namely:

- trial and error/comparing results
- reverse engineering
- based on sensitivity and precision
- depends on the study design.

Apart from adding a subject search, do you amend methodological search filters?

The questionnaire sought to find out whether or not searchers amend filters. Four respondents (4.4%; 4/90) indicated that they always amend search filters. Over half of the respondents said that they sometimes amended filters (55.6%; 50/90) and one-third indicated that they do not make changes to filters (33.3%; 30/90) (*Table 31*).

Why, typically, do you amend search filters?

Twenty-six (28.9%) out of 90 respondents indicated that they amended filters to improve sensitivity and/or specificity, for example:

> *We are afraid to miss things so we amend filters to enhance sensitivity.*
>
> *Sometimes to make them a little shorter or to increase/decrease sensitivity.*

TABLE 30 How do respondents decide which filter to use?

Typical practice	Number of respondents	Percentage of respondents
I research the available filters and choose the best for my purposes	51	56.7
I use the filters available in the database interfaces that I use, e.g. Clinical Queries	48	53.3
Custom and practice – I've always used the same filters	34	37.8
Guidance from a colleague	34	37.8
I follow standard operating procedures/guidance on filters provided by my organisation	22	24.4
I use international/national guidance on best practice	21	23.3

Reproduced with permission from Beale *et al.*[132] © 2014 The authors. Health Information and Libraries Journal © 2014 Health Libraries Journal. *Health Information & Libraries Journal*, **31**, pp. 133–147.

TABLE 31 Frequency with which respondents amend search filters

Frequency	Number of respondents	Percentage of respondents
Always	4	4.4
Sometimes	50	55.6
No	30	33.3
Did not respond	6	6.7
Total	90	100.0

Where there are inappropriate results returned I may be able to improve specificity.

Either to broaden or narrow the scope of a search.

How do you amend search filters?

Twenty-eight (31.1%) out of 90 respondents indicated that they amended search filters by adding or removing terms. Other forms or methods of amendment reported by respondents were:

- adapting to another database
- looking at adjacency or truncation
- researching MeSH terms and adding free text
- examining which lines of syntax are producing zero or too many results
- by inclusion of keywords and weighting word algorithms
- based on advice from other librarians.

Do you test and document the effects of any amendments you make?

All who responded to this question indicated that they always or sometimes amend search filters and, of these, a majority (83.3%; 45/54) also indicated that they tested the effects of the amendment (*Table 32*).

Respondents reported that they test the effects of any amendments by:

- 'eyeballing' results
- conducting a 'before and after' comparison
- assessing whether or not key relevant articles have been identified.

About three-quarters of the respondents (75.9%; 41/54) who make changes to search filters document the changes that they make (*Table 33*).

TABLE 32 Number and percentage of respondents who test the effect of search filter amendments

Do you test the effect of search filter amendments?	Number of respondents ($n = 54$)	Percentage of respondents
Yes	45	83.3
No	9	16.7
Total	54	100.0

TABLE 33 Number and percentage of respondents who document the amendments to search filters when they write up their searches

Do you document search filter amendments?	Number of respondents ($n = 54$)	Percentage of respondents
Yes	41	75.9
No	11	20.4
No response	2	3.7
Total	54	100.0

A wide range of approaches was reported for documenting amendments to search filters, with about three-quarters of respondents indicating that they comprehensively documented changes. Some examples of the broad nature of responses to this question are shown in the following quotations:

> *Usually reproduce entire search string and provide written summary of rationale for changes and effects.*
>
> *I keep spreadsheets of search terms where each column is an iteration of my search, with notes on why changes occur so that I have a record and a rationale.*
>
> *I make a note of where I adapted the strategy from and then save the search strategy in a word document and also in the database where possible.*
>
> *Narrative included in both the methods section of the review and in the annexe.*
>
> *I record that the filter has been adapted for use in other databases.*
>
> *I add some comments to the search line.*
>
> *Save the searches for future reference but how they are written up depends on client requirements.*
>
> *Only to the degree that I may save the search in my saved search file . . . and save the search to our search recording software.*
>
> *Yes, but not always!*

Keeping up to date

Respondents were asked to select from a list the method(s) that they use to keep up to date with search filters (*Table 34*). The most frequently reported method of keeping up to date was through professional development meetings and training events (74.4%; 67/90).

Over 60% of respondents reported keeping up to date by the following methods:

- reading journal articles
- reading e-mail lists
- through information provided by managers/work colleagues.

TABLE 34 Methods of keeping up-to-date

Method of keeping up to date	Number (%) of respondents answering ($N = 90$)		
	Yes	No	No response
Professional development meetings and training events	67 (74.4)	14 (15.6)	9 (10.0)
Reading journal articles	60 (66.7)	21 (23.3)	9 (10.0)
Information provided by managers/work colleagues	60 (66.7)	19 (21.1)	11 (12.2)
E-mail lists	57 (63.3)	23 (25.6)	10 (11.1)
Websites	50 (55.6)	28 (31.1)	12 (13.3)
Current awareness services	24 (26.7)	50 (55.6)	16 (17.8)
RSS feeds	9 (10.0)	64 (71.1)	17 (18.9)

RSS, really simple syndication.

Respondents were asked to indicate which specific current awareness resources they used to keep up to date. Some respondents indicated that they used more than one resource. Current awareness searches set up in databases such as MEDLINE and EMBASE were the most frequently cited approaches, followed by tables of contents services. The resources cited were:

- database alerts/current awareness searches (nine respondents)
- tables of contents services (six respondents)
- e-mail discussion lists (two respondents)
- really simple syndication (RSS) feeds (two respondents)
- American College of Physicians journals (otherwise unspecified) (one respondent)
- AETMIS (Agence d'évaluation des technologies et des modes d'intervention en santé) current awareness service (one respondent)
- AHRQ current awareness service (one respondent)
- Cochrane (otherwise unspecified) (one respondent)
- discussion with colleagues (one respondent)
- end-of-life care (otherwise unspecified) (one respondent)
- library blogs [including Krafty Librarian, iLibrarian, Phil Bradley, OCLC, ScienceRoll, MedScape] (one respondent)
- NICE internal current awareness bulletin (one respondent)
- Palliative Care Journal Club (one respondent)
- WebSite-Watcher e-mail alerts (one respondent).

Respondents were asked to indicate which websites they used to keep up to date. Some respondents provided more than one resource. The ISSG Search Filters Resource was the most frequently cited website (25.6%; 23/90). The SIGN, McMaster Hedges team, MEDLINE and Cochrane resources were also mentioned by between four and nine respondents. As previously, some resources were described using various names and some assumptions have been made about groupings. The websites cited were:

- InterTASC/ISSG/CRD (23 respondents)
- SIGN (nine respondents)
- McMaster Hedges team (seven respondents)
- MEDLINE/PubMed/US NLM (National Library of Medicine) (otherwise unspecified) (five respondents)
- Cochrane (Collaboration/IRMG/Handbook/Library) (four respondents)
- BMJ Clinical Evidence (three respondents)
- Cindy Smith's blogspot (three respondents)
- University of British Columbia Library Health Library Wiki (two respondents)

- BestBETs (one respondent)
- Centre for Evidence-Based Medicine (one respondent)
- Google (one respondent)
- government and non-governmental health organisations (one respondent)
- HTAi Vortal (one respondent)
- Knowledge Network – shared space for Scottish librarians (one respondent)
- national and university websites (otherwise unspecified) (one respondent)
- US NLM e-text on HTA (one respondent)
- Ovid databases (otherwise unspecified) (one respondent)
- World Health Organization (one respondent).

Respondents were asked to indicate which e-mail lists they used to keep up to date. The most frequently reported lists were Cochrane e-mail discussion lists (19 respondents) and expertsearching (nine respondents). National medical librarian discussion lists for the USA, Canada and the UK were frequently mentioned. The e-mail lists cited were:

- Cochrane information specialist e-mail discussion lists (IRMG/librarians/methods/TSCs) (19 respondents)
- expertsearching (nine respondents)
- CANMEDLIB (eight respondents)
- LIS-MEDICAL (eight respondents)
- MEDLIB-L (eight respondents)
- HTAi Information Resources Group (seven respondents)
- Evidence-Based-Health (five respondents)
- InterTASC ISSG (five respondents)
- CLIN-LIB (two respondents)
- local health libraries network (unspecified) (two respondents)
- EAHIL-L (one respondent)
- Evidence-Based-Libraries (one respondent)
- Health Sciences Libraries Group discussion group for the Health Sciences Libraries Group of Ireland (one respondent)
- LIB-HELIX discussion group for library staff in the NHS South Central area of the UK (one respondent)
- LIS-NURSING (one respondent)
- medical librarian lists (unspecified) (one respondent)
- NCC-information specialists (one respondent)
- professional e-mail lists (unspecified) (one respondent)
- SYS-REVIEW (one respondent)
- WEBENZ e-mail list for medical information specialists in the Netherlands (one respondent).

Respondents were asked to indicate which RSS feeds they subscribed to to help to keep up to date. RSS feed usage was low, with the most frequently reported feed being Cindy Smith's (most likely Schmidt's) blogspot (http://pubmedsearches.blogspot.ca/) and Evidence Based Library and Information Practice (EBLIP).[134] The RSS feeds cited were:

- Cindy Smith's (most likely Schmidt's) blogspot (three respondents)
- EBLIP (three respondents)
- PubMed New and Noteworthy/PubMed Technical Bulletin (two respondents)
- *Health Information and Libraries Journal* (one respondent)
- journal tables of contents (unspecified) (one respondent)
- librarian blogs (unspecified) (one respondent)
- LISNews (one respondent)
- medical libraries (unspecified) (one respondent)
- OvidSP Updates (one respondent)
- websites (unspecified) (one respondent).

Respondents were also asked to provide other methods (not listed above) that they used to help keep up to date. The methods reported included:

Check my file of papers on search filters.

Search for filters when one is required (PubMed or Google or post a query to MEDLIB-L or CANMEDLIB).

If there were changes in the Cochrane Handbook I would incorporate these into my work.

Meetings of the Cochrane Information Retrieval [Methods] Group.

My colleagues and I have a journal club where articles such as these are often chosen.

Other colleagues within my unit.

Attend workshops.

If you have had to choose between methodological search filters, what features or information has helped you to do so?

Respondents were asked how, when faced with a choice of methodological filters, they chose a filter. Several respondents (16.7%, 15/90) said that they required information on the performance of a filter (sensitivity, specificity and precision), with other respondents (11.1%; 10/90) requiring published reports and evaluations of the filter. Five respondents (5.6%, 5/90) required information on authorship and five looked to colleagues for advice.

Other approaches reported by respondents included:

Personal knowledge and testing.

Length – the shorter the better.

Relevant database.

Focus.

Flexibility/modifiability.

Testing.

InterTASC site/ISSG.

Choose the ones that look logical based on my experience.

Search words used.

If you report your search process do you describe the filters that you have used?

Most respondents (86.7% 78/90) reported that they described the search filters that they used, with 4.4% (4/90) of respondents reporting that they did not (*Table 35*).

If you report your search process do you justify your choice of filters used?

Just over half of respondents (57.8%; 52/90) reported that they did not justify their choice of filter when writing up their search, whereas approximately one-third (32.2%, 29/90) of respondents reported that they do provide a justification (*Table 36*).

TABLE 35 Number and percentage of respondents who provide a description of the search filters used

Do you describe the filters used in the search process report?	Number of respondents ($n = 90$)	Percentage of respondents
Yes	78	86.7
No	4	4.4
No response	8	8.9
Total	90	100.0

TABLE 36 Number and percentage of respondents who provide a justification for the search filters used

Do you justify your choice of search filters used?	Number of respondents ($n = 90$)	Percentage of respondents
Yes	29	32.2
No	52	57.8
No response	9	10.0
Total	90	100.0

What do you think are the benefits of using methodological search filters?

The most frequently reported benefits of using methodological search filters were that they helped to focus results (42.2%; 38/90), they are tried and tested (18.9%; 17/90), they save time (10%; 9/90) and they offer transparency and consistency (5.6%; 5/90).

Respondents also reported other benefits, including:

> *help estimate workload in project planning.*
>
> *[to enable] conceptual mapping of thoughts.*
>
> *rerunning is easy, results are comparable.*

What do you think are the limitations of using methodological search filters?

Respondents reported that the most frequent concerns they had about using a methodological search filter were that studies would be missed (37.8%; 34/90), filters were not always fit for purpose (22.2%; 20/90), filters lacked transparency or were hard to appraise (10%; 9/90) and filters were reliant both on the competence of the filter developer and on the adequacy of record indexing (14.4%; 13/90).

Other limitations included:

- can sometimes be hard to choose between filters (one respondent)
- lack of instructions for publishing (one respondent)
- sometimes hard to explain filters to researchers (one respondent)
- the 'perfect filter' is not always available and so 'the next best thing' is used, which is not ideal (one respondent)
- too many results (one respondent).

What information would help you to choose which filter to use?

Respondents reported that they would like information on filter performance measures such as validation (27.8%; 25/90) and sensitivity and specificity (20%; 18/90), and a description of the filter (16.7%; 15/90).

Other information requirements included:

- results of own testing (11 respondents)
- colleague recommendations/discussion (six respondents)
- the database (four respondents)
- knowledge of the creator/developer (three respondents)
- simplicity/understandability (three respondents)
- ease of use (including automatic loading) (two respondents).

Respondents reported that the main factors that would make choosing a filter easier were the availability of a critical appraisal or evaluation (17.8%; 16/90) and more information (such as the effectiveness of the filter, what it does/provides, what it excludes, its limitations, when it was last updated; advantages and disadvantages; sensitivity and specificity; how it has been tested) (16.7%; 15/90). Respondents also reported that they wanted to be confident in the author/developer (11.1%; 10/90).

Other factors cited as making it easier to choose which filter to use were:

- the presence of a central storage location (seven respondents)
- better expression/presentation of results (four respondents)
- greater consistency in the methods used (one respondent)
- availability/accessibility in all databases (problem with CINAHL on EBSCO*host*) (one respondent)
- better labelling/indexing of articles so that they might be more easily retrieved (one respondent)
- more up-to-date coverage on the CRD (i.e. ISSG Search Filters Resource) website (one respondent)
- more 'professional noise' about a new filter (one respondent)
- the availability of synopses of filters (one respondent).

What methodological search filters would be useful to you?

The respondents had a wide range of requirements for new filters:

- economic/economic evaluation/cost–benefit/cost–utility studies (five respondents)
- all research/study designs (in one filter) (two respondents)
- controlled trials/controlled studies in the public health field (two respondents)
- diagnosis/diagnostic studies (two respondents)
- a combination of RCTs and systematic reviews/meta-analysis in one filter
- aetiological studies
- burden of illness studies
- case–control studies
- case series
- clinical audits
- clinical trials
- cross-sectional studies
- epidemiological studies
- full-text searches
- guidelines
- HTAs
- interrupted time series
- meta-analyses
- non-RCTs
- observational studies
- process evaluations
- qualitative studies
- quasi-experimental studies
- RCTs

- social sciences methodologies (other)
- specific methodologies
- systematic reviews.

Respondents also had requirements for filters capturing other issues:

- adverse effects/events/harms
- age groups
- children/paediatrics
- demography
- disease specific/technology specific
- emergency departments
- errata
- hospital management (non-clinical)
- hospital setting
- magnetic resonance imaging
- older people
- patient-centred outcomes
- patient experience
- prognosis/prognostic studies
- programmes and services
- public health, especially health protection and infection control
- retracted or withdrawn articles
- therapy.

With respect to databases, respondents expressed an interest in the following database-specific filters:

- a more precise RCT filter for EMBASE
- a validated/Cochrane-recommended RCT filter for EMBASE and other databases, for example CINAHL
- a definitive filter per methodology per database
- Education Resources Information Center (ERIC) filters
- filters validated for more databases than EMBASE and MEDLINE.

Other comments about filters were invited and the following were noted:

> *Clearer ones.*
>
> *Please, no further filters – more work on limitations of filters and dissemination work on alternative/better methods is necessary.*
>
> *UK studies.*

Further observations on methodological search filters as a tool for information retrieval

Respondents provided additional views on methodological search filters as a tool for information retrieval, which have been grouped in the following sections under limitations and benefits.

Limitations

> *As an 'ordinary searcher' I find the choice in Hedges totally bewildering.*
>
> *From a clinical point of view the whole business – if well intentioned – seems fraught with difficulty and uncertain relevance.*

Because I haven't really understood them when I've looked, I tend to avoid them. . . . Part of the problem is I think it varies depending on which interface you use to search a database.

Much effort to produce these but often used alongside other dubious practices, e.g. discarding papers that have no abstracts (SIGN, in particular, do this) so making their precision rather pointless.

Ultimately, even after long discussions with clients, I have to change them!

There is too much reliance on the present filters: they almost have a golden status which means it is objectively difficult to manoeuvre away from one or other or make a tweak. Non-IS [information specialist] types get nervous believing in the infallibility of written filter.

Even though it is great to have a site such as InterTASC, it is very difficult to locate actual filters on it.

I love the built in hedges to databases that makes it easy to use 'click a button' but at the same time, I like to have what's 'under the hood' easily available to look at or in order to report your own searches.

. . . please accept search filters are out-of-date methodology. Please don't develop more filters! That's not helpful for high-quality information retrieval in systematic review, HTA and guideline-writing! Information technology development offers better methods in 2011 than in the early 1990s.

Benefits

They are necessary for the work of evidence-based health care – future development should focus on maximising the precision of filters.

What might be more useful [than methodological filters] are more topical filters as is being started with the MedTerm Search Assist Database (http://www.hsls.pitt.edu/terms/).

I am very grateful to the people who have developed validated filters.

When I started to work here, I learned that filters were not used except for 'quick and dirty' searches. The majority of our searches are for reviews, so I learned not to use filters and accept rather high numbers of references. Gradually, I am reconsidering this as time is an important factor too and good filters could save much time!

I like to inform researchers, students, research co-ordinators, doctors, nurses etc. about these filters as I think they will be very helpful for them.

Researchers despise going through 20,000 articles so we need to find ways to make precise searches without losing too much sensitivity.

Is there a central repository with features for comparison?

Discussion

In 2004, Jenkins and Johnson[5] reported that, although researchers were aware of filters, there was a low level of usage. Since then it appears that more people are using filters to inform their research and filters are being used for a range of searching tasks.

The questionnaire reported in this chapter has several limitations. Although we do not know what proportion of search filter users we reached, questionnaire analyses showed that our sample included

librarians and other information specialists and researchers involved in supporting systematic reviews, technology appraisals and guideline development, all of whom represent our target audience. From the e-mail lists that respondents reported being members of, we can tell that many were information specialists supporting the production of HTAs, guidelines and systematic reviews. We therefore expect the results to be broadly generalisable to such librarians and other researchers. The e-mail lists that we sent the questionnaire to had at least 2857 subscribers, but we do not know how many people, in total, the e-mails reached because people may have been members of multiple lists. In addition, respondents had other ways to find out about the survey, such as Twitter. Moreover, the survey invitation was sent to general health-care librarian lists as well as specialised lists and many members of the former would not have been competent to respond to the survey. The e-mail lists we used, however, ranged from lists with high proportions of information specialists, with roles similar to those of NICE information specialists, to more general lists, whose members might not routinely use search filters.

The questionnaire that we developed was quite lengthy and, in retrospect, might have benefited from being shorter. The response rate, however, to early questions was similar to that to later questions, suggesting that the length of the questionnaire did not act as a deterrent to any of the individuals who actually submitted a response. It might have helped to achieve more standardised results and fewer ambiguous answers if we had given respondents more multiple choice questions. Respondents described resources quite vaguely at times and sometimes the same resource was described using several different names. We have made some assumptions about the variant naming of widely used resources to provide a more succinct report and to ensure that the most frequently reported resources are identified as such. We have not, however, routinely corrected what may be 'errors' in the responses, for example when certain filters may be incorrectly described or ascribed to the wrong author or organisation.

When do searchers and researchers use search filters?

The awareness and use of search filters seems to have developed considerably in the decade since the publication of the article by Jenkins and Johnson.[5] Most respondents seem to know where to look for filters from well-established producers and collections. The responses, however, demonstrate a wide variation in the confidence with which questionnaire respondents choose filters. There are also contradictions between the difficulties that respondents express in terms of selecting between filters (acknowledging the possible complexities of filter design) and the commonplace practice of searchers adapting published strategies to fit their own requirements (ignoring the fact that many filters are designed to perform in a quite specific way). Several respondents have developed their own filters for local use. The responses indicate that search filters are used more frequently for large-scale reviews and slightly less often for simpler scoping and rapid searches. This may reflect different practices in scoping and rapid searches because fewer resources will be searched and less sensitive subject searches will be employed because of the limited timescale. Adding a filter to an already focused search might be seen as risking missing studies. For all types of searches, search filters offer an opportunity to focus the numbers of records retrieved, which can be helpful when time is limited. Search filters are predominantly viewed by respondents as a tool to maximise sensitivity rather than precision (although this is not the intended objective of all filters), but seem to be used to achieve optimal sensitivity and precision.

What information would help researchers choose between filters?

The responses to the questionnaire have many messages for search filter designers. Filter performance measures need to be signposted more clearly and succinctly to help searchers make better use of the available filters. Filter and website designers should present less information (to avoid information overload) and ensure that performance information can be clearly seen. Respondents also reported that they wanted to be confident in the author/developer. While the provenance of filters is clearly important to some searchers, there are no established parameters to measure this confidence. Clear authorship labelling and the provision of detailed methods to show the robustness of the development methods would not only assist users of filters but also help filter designers achieve recognition for their filters. The convenience of having filters by well-established producers available within database interfaces (such as the PubMed Clinical Queries filters) encourages their use. The most convenient search filters, however, may not always

be the best for particular tasks and searchers and researchers need to know how to choose when a range of sensitive, precise or 'optimal' strategies is offered. Respondents require more information on the validation of search filters. They value and use resources such as the ISSG Search Filters Resource and the filters of the McMaster Hedges team. The former provides a list of all identified methodological search filters in one place, by study design and by database, which has a convenience factor. The latter provides search filters developed using documented methods within database interfaces, with filters 'badged' with the authority of both the research team and the US NLM. In contrast to the methodological and publisher quality seals of the McMaster filters, the BMJ Clinical Evidence and SIGN websites provide little information on filter production and/or validation. The filters on these websites, however, seem to be widely used, suggesting that authorship is the seal of quality.

Respondents did not necessarily feel that all of their requirements were currently being met. They would like translations of filters for different databases and interfaces, more strategies independent of indexing language (to facilitate transferral across databases) and filters for a wider range of study designs and other topics. This provides a research agenda for any search filter authors willing to take up the challenge.

Respondents keep informed about developments in search filters through a wide variety of methods and resources, which suggests that search filter and website designers face a marketing challenge. Highlighting new filters to key audiences such as information specialists and systematic reviewers by inclusion in resources such as the *Cochrane Handbook*[133] and the ISSG Search Filters Resource[6] would help to promote new filters beyond the simple publication of a journal article. In addition, a large number of e-mail lists are used for current awareness purposes, and the promotion of new filters through these lists would seem to be an efficient way to reach potential users.

Although the use of search filters seems to be quite widely documented and amendments are noted in search reports, there seems to be scope for promoting clarity around the use and amendment of search filters. This, again, is an issue for filter authors and website producers. There is clearly a large amount of ad hoc filter amendment work being undertaken: searchers take filters and adapt them for their own purposes. This would seem to indicate a lack of awareness that the filters may be designed for a purpose or have been arrived at after extensive exploration (increasingly using textual analysis techniques) to justify the use of specific terms and the absence of others. The performance assessment of amended search filters does not seem to be a priority for many searchers. Filter developers should consider how they want their filters to be used and perhaps attach guidance or caveats to the filters. Guidance for filter adaptation may also be merited so that filter developers are credited for the original work but absolved from the effects of the adaptations. Many filter developers retain their gold or reference standards and might be willing to test adaptations.

The original impetus for many search filters was to maximise sensitivity but, increasingly, possibly because of limited resources, searchers seem to be demanding improvements in precision. Future filter developments (for interfaces that use Boolean searching) need to continue to improve precision while maintaining sensitivity. The advent of full-text searching and semantic analysis of both full-text and bibliographic records may see filters used in different ways in the future. For example, sensitive filters might be used to identify records from databases and these results might then be processed using semantic analysis software trained to identify records of specific types. The results could then be used to revise the search filter and improve the precision of the search results. This approach will have search algorithms (filters) that are more like semantic rules than the dichotomous (relevant/not relevant) search filters that we see used in bibliographic databases such as MEDLINE. Textual analysis approaches have been used in the design of searches.[22,135] The extent to which textual analysis alone can be relied on in the future to distinguish relevant records from irrelevant records is under investigation.[59,135] When using semantic analysis approaches the onus will be on the searcher to select the performance levels, that is, to choose an acceptable probability of a record being relevant.

Conclusion

Search filters are used mainly for reducing the size of large result sets (introducing focus) and assisting with searches that are focused on a single study type. Searchers use several key resources to identify search filters but may find choosing between filters problematic. Features that would help with filter choice include making information about filters less technical, offering ratings and providing more detail about filter validation strategies and filter provenance.

Chapter 5 Suggested approach to measuring search filter performance

Introduction

This chapter outlines a suggested approach to test the retrieval performance of search filters, with a view to encouraging searchers to contribute to the larger picture of search filter performance. Once piloted, this approach could form the basis of published guidance on how to conduct search filter performance testing. Recommendations are based on the findings of the reviews, interviews and questionnaire, published literature and the cumulative experience in search filter research of the authors.

Search filter studies, to identify studies that use specific research designs (such as RCTs), which are purposefully developed and published in journal articles, typically present two or three measures of performance. These measures tend to be based on testing filters on one or two sets of relevant records (known as reference sets/gold standards). Our research has shown that the performance of filters across different disciplines, questions and health databases is largely unknown (review B) and that a range of different performance measures is reported in articles describing search filters (review A). There is a paucity of published data on how searchers select filters (review D) although, when questioned, experienced searchers described informal and pragmatic experimenting with filters or relying on the provenance or published performance measures to aid selection (interviews). In addition, respondents to the questionnaire mentioned using filters that are available in the database interface, consulting colleagues or having filters that they always use.

Both interviewees and questionnaire respondents expressed a desire for the performance measures of published filters to be signposted more clearly and succinctly. Data on the performance of filters in different reference standards are needed to help searchers to assess whether or not filters perform consistently and also to detect topics or fields in which the performance of a filter may be better or worse. Collecting performance data and sharing them through a central resource (such as a website) would mean that there is greater availability of information for all users of search filters. The approach proposed here offers ways to collect search filter performance data and to report them on the ISSG Search Filters Resource website [see https://sites.google.com/a/york.ac.uk/issg-search-filters-resource/home (accessed 22 August 2017)].

Examples are provided to show that search filter performance measurement can be conducted as part of systematic reviews or other projects involving extensive searches.

Measuring search filter performance

There are several aspects to measuring search filter performance:

- Which performance characteristics should be measured (e.g. sensitivity, precision)?
- How should a performance measure be ascertained (e.g. how to develop a reference set)?
- How can performance measurement be carried out most efficiently?

Which performance characteristics should be measured?

When considering the measures that are most useful to users of search filters, the following are recommended based on the responses to the interviews and the questionnaire survey carried out to inform this research. These measures are also those most frequently reported in the literature (reviews A and B):

- sensitivity
- precision or NNR.

Sensitivity is defined as the number of records in the reference set that are retrieved by a search filter as a proportion of the total number of records in the reference set. It is therefore a crucial performance issue for many searchers, especially within the context of many systematic reviews and HTAs, in which searchers are usually focused on retrieving as much relevant evidence as possible. This may be less of a concern in reviews of qualitative evidence.[136]

Precision is defined as the number of reference set records retrieved by a search filter as a proportion of the total number of records (relevant and irrelevant) retrieved. It is also a crucial issue for searchers involved in evidence synthesis because, in seeking to achieve high sensitivity, retrieval rates are often high and the precision tends to be low. One study reports that 2–3% precision is typical of searches undertaken in systematic reviews[137] but experience suggests that precision is often much lower than that. Precision is also a concept that is of relative importance to searchers as they are likely to be more tolerant of low precision when low numbers of records are retrieved (e.g. in a search topic when there are few research reports) than when high numbers of records are retrieved (e.g. in a search of breast cancer). NNR offers a precision-based metric to indicate the workload involved in identifying relevant records when using a specific filter.

An additional performance measure could be collected that might assist with estimating workload. We have called this 'reduction in number needed to read' (see *Glossary* and also review C and Whiting *et al.*[2]). This indicates how far adding a filter to a subject search will reduce the workload involved in processing records by showing the reduction in the number of records that will need to be screened. A small reduction in the number needed to screen may indicate that using a filter is not helpful in reducing the workload involved in assessing retrieved records for relevance, whereas a large reduction in number needed to screen may encourage the use of a filter.

How should a performance measure be ascertained?

If sensitivity and precision are the focus of performance measurement, the following issues are crucial for robust measures:

- having a definition of the criteria for building a reference set
- having a reference set of relevant records (to measure sensitivity and precision/NNR)
- having a results set containing all records retrieved by hand-searching or records retrieved by RR methods or the total number of records retrieved by a search of a database using a subject search strategy (to measure precision/NNR)
- having search filters that are suitable for the database interface being used to search for records or that have been translated carefully to be used in another database interface.

These issues are discussed in more detail in the following sections.

Reference set criteria

To build a reference set of relevant records, the inclusion criteria for a record to be assessed as relevant to the reference set need to be described in adequate detail. The inclusion criteria may include definitions of a population, an intervention or an outcome or other features against which a record can be assessed for relevance. An example of a description of a reference set is shown in *Box 1*. The descriptions are important to ensure that the reference set includes the same types of studies that the filter being tested is designed to retrieve and should be as detailed as possible.

Identifying a reference set of relevant records

The reference set should be representative of all relevant records to minimise bias and increase the robustness of the results and should be large enough to provide reliable results (review C). As reviews A and B demonstrated, there are two widely used methods of identifying a reference set of relevant records (and probably many variants):

1. hand-searching database records or sets of publications (usually journals) to identify all of the records that meet a set of explicit criteria[69,133]
2. using the RR technique to create a reference set based on the results of a systematic review.[52]

BOX 1 Example description of a reference set

The reference set includes records that meet the following criteria:

- reports of RCTs (trials with two or more arms in which patients are allocated to an arm using a randomisation method; the trial may or may not be blinded)
- population – women aged ≥ 65 years
- condition – experiencing urinary incontinence
- outcomes – reporting impact on quality of life
- intervention – low-caffeine and low-sugar drinks compared with caffeinated drinks (low or high sugar).

Other more subjective methods of creating a reference set, such as using personal collections of records, are not usually recommended. This is because the methods used to create the reference set from personal collections may mean that the records are not generalisable to the records that a filter is aiming to identify. The methods are also unlikely to be transparent and replicable, may be hard to characterise by factors such as date and may be difficult to report clearly.

Hand-searching

Hand-searching can be conducted in various ways[138] and the methods used to identify a set of publications (books, records, conference proceedings or journals) to hand-search should be clearly reported. Methods used to identify a set of publications include selecting a random sample of database records or selecting journals to hand-search based on a frequency analysis of documents in which relevant records appear.[139] For the former method, it may be necessary to assess the sample size necessary to assume a representative sample of records (review C).[83] One way to do this would be to carry out a series of searches to establish the proportion of studies with the required design in the database and then calculate the required sample size.[47] The choice of timespan over which relevant records are published should also be considered: 1 year may not capture potential changes in terminology or reporting developments. It may be best to optimise the usefulness of the reference set by searching a range of years as well as a range of documents relevant to the filter and the database coverage. It may also be advisable to search both subject-specific as well as more general journals (as reported in review C).

Any limitations in terms of the generalisability of the selected publications to all similar publications should be made clear. The identification of a results set of (database) records to be assessed for relevance may be achieved by searching using a general (high-level) indexing term (as reported in review B).

It should be acknowledged that developing a reference set using hand-searching can be time-consuming, especially as, ideally, developing a hand-searched reference set should be conducted by at least two independent assessors to minimise selection bias.

Relative recall

The RR approach to identifying a reference set of publications should usually be less resource intensive than hand-searching, but does require a critical assessment of the searching used in the underlying review.

Using the RR technique to create a reference set of publications, based on the results of a systematic review, has been described in detail by Sampson *et al.*[52] RR has been used to develop reference sets for testing search filter performance by a number of researchers (reviews A and B).[2,48,49] The studies included in a systematic review (or other research project), in which extensive searching using sensitive search strategies and other approaches to study identification have been employed, are taken as a quasi-reference set. The assumption is that the exhaustive search has approached the identification of all relevant studies.

The quality of the RR reference set relies on the extensiveness of the search, the adequacy of the subject search strategies used to identify studies and the presence of clear relevance criteria for the selection of records. The criteria used to select studies, however, cannot always be translated to the search strategy.

For example, a sample size minimum will exclude small RCTs from the reference set but sample size cannot be readily incorporated into the search filter. This will result in artificially reducing the precision of the search filter as small RCTs have been excluded from the reference set although they meet the purpose of the filter in identifying all RCTs. It is preferable that the search terms used in the original review search strategies do not include any of the methodological search terms included in the filter being tested, as this can lead to bias by artificially inflating the sensitivity of tested terms.

The subject search contains terms designed to capture a specific topic such as an intervention in a disease or an outcome following an intervention. It should be assessed in terms of its ability to adequately find relevant records, that is, records that address the search question. The search question is the research topic that the search has been designed to answer through the capture of relevant records. The strategy should be checked to ensure that the appropriate index terms and a suitable range of free-text terms have been used with the correct use of Boolean operators, truncation and proximity operators. If a subject search is more precise than the search question, then sensitivity may be compromised and precision is likely to be maximised. For example, if the search question relates to breast cancer and the subject search focuses on stage IV breast cancer then the subject search is less sensitive than the search question. If the search strategy is more sensitive than the search question, the filter precision may be compromised unfairly. Continuing the example, if the subject search is constructed to look for cancer records, then it will be far more sensitive and less precise than the search question. The adequacy of the strategy should be assessed using the Peer Review of Electronic Search Strategies (PRESS) checklist.[140,141] If the search strategy is judged to be inappropriate or inadequate for the search question, it may be better to select another review for testing. The subject search may have specific exclusions, such as animal studies, and the impact of explicit exclusions on the results should be considered.

If adaptations need to be made to the subject search (perhaps because the search was developed for a different interface to the database), these adaptations should be made carefully and should be reported in detail, with an assessment of how far they differ from the original search.

Relative recall has the benefit that it is a relatively straightforward and economical method of identifying a reference set at the same time as undertaking a review. The reference set will, however, tend to be highly specific and confounded by the subject searches undertaken to populate the research project. Using multiple reviews has been suggested to increase the robustness of the reference set.[52]

The RR reference set will have been created at a specific point in time; the same subject search run subsequently will find more results and it is difficult to recreate the status of a database at a specific point in time. Methods to remove later studies, to approximate to the state of the database at the time of the original searches, may be used but should be documented. One such approach might be to remove all records with database entry dates later than when the original search was undertaken.

Creating the reference set for testing

To create a reference set for testing the performance of a search filter, the relevant records identified from hand-searching or from a systematic review or reviews have to also be identified in a specific database using a known-item search approach, such as searching by author name or title. The records are then combined to create the reference set. The search filter can then be run in the database and the number of records it retrieves from the reference set available within that database can be ascertained.

The results set

The results set can be variously defined. It may be the total number of records that are retrieved by hand-searching, the total number of records retrieved by RR methods or the total number of records retrieved by a search of a database using a search strategy.

For testing the performance of a filter in retrieving records from a reference set identified by hand-searching publications (including database records), the results set must include all records in the database segment or publications searched (both relevant and irrelevant).

For testing filter retrieval of a RR reference set, the results set consists of the records retrieved by the subject search used in the systematic review for the database being searched.

Search filters

The results of the interviews and questionnaire suggest that experienced searchers consult a variety of sources to identify search filters. The most frequently mentioned was the ISSG Search Filters Resource,[6] which was also the source used for reviews A and B. This collaborative venture identifies and collates a wide range of methodological search filters, organised by study design and by database.[6]

Searchers are likely to look closely at the trade-off between sensitivity and precision/NNR when deciding which methodological filters to use to match the purpose of a search, for example high sensitivity for a comprehensive search or higher precision for a scoping search. The choice of a filter may also need to take into account other factors to check transferability to the intended database:

- The sensitivity and precision of the subject search.
- The characteristics of the intended database, such as the indexing practices, facilities and search options (e.g. proximity operators) available, which will determine suitability for translation into other databases and/or other service providers.
- Variations in reporting and consistency of study methods between the subject areas of the intended search and the filter reference set.
- Variations in the ways that authors define their study designs in abstracts should be accommodated by the filter.
- The currency of the filter. Subsequent changes in database indexing from when the filter was created will determine suitability and the need for adaptation.

The search filters to be tested should be used as intended by the authors. For example, a sensitivity-maximising filter to identify reports of RCTs in MEDLINE designed using the OvidSP interface should really be tested for that purpose. The filter should be obtained from the original publication to ensure accurate use (filters can sometimes be changed or unintentionally mistyped when used and reported by other authors). However, if the filter needs to be translated to another database and/or interface it should be translated carefully. The original and translated filters should be reported, along with an assessment of the impact of any changes on retrieval performance. An example of a translated filter is provided in *Table 37*.

TABLE 37 Example of an original and translated filter

Original Ovid strategy	Translation to PubMed	Notes
exp "Sensitivity and Specificity"/	"sensitivity and specificity"[mh]	PubMed explodes by default
sensitivity.tw.	Sensitivity [tiab]	Used [tiab] to restrict to title and abstract
specificity.tw.	Specificity [tiab]	Used [tiab] to restrict to title and abstract
((pre-test or pretest) adj probability).tw.	"pre-test probability"[tiab] OR "pretest probability"[tiab]	PubMed has no proximity operators so we have used the phrase option. This only works, however, if these phrases are predefined by the US NLM. We could also try a search using AND, although this is much more sensitive than the original: (pre-test [tiab] AND probability[tiab]) OR (pretest [tiab] AND probability [tiab])
post-test probability.tw.	"post-test probability" [tiab]	The same issue about proximity operators applies. In addition, the original search does not compensate for non-hyphenation in this line, whereas it did in the previous line (i.e. posttest is not searched). We have also omitted the 'posttest' option to ensure that we do not introduce additional differences
predictive value$.tw.	"predictive value*" [tiab]	Used [tiab] to restrict to title and abstract
likelihood ratio$.tw.	"likelihood ratio*" [tiab]	Used [tiab] to restrict to title and abstract
or/1–7	#1 OR #2 OR #3 OR #4 OR #5 OR #6 OR #7	

How can performance measurement be carried out most efficiently?

A flow diagram showing the key steps in conducting search filter performance measurement using a hand-searched reference set is shown in *Figure 11*. The process should be fully documented as it is undertaken and ideally the search filter should be tested on its own, without the addition of a subject filter, as the hand-searched documents provide the test bed. The question of how far the hand-searched documents are representative of all documents that might yield relevant records should be discussed.

A flow diagram showing the key steps in conducting search filter performance measurement using a RR reference set is shown in *Figure 12*. As above, the process should be fully documented as it is undertaken and the same caveats apply.

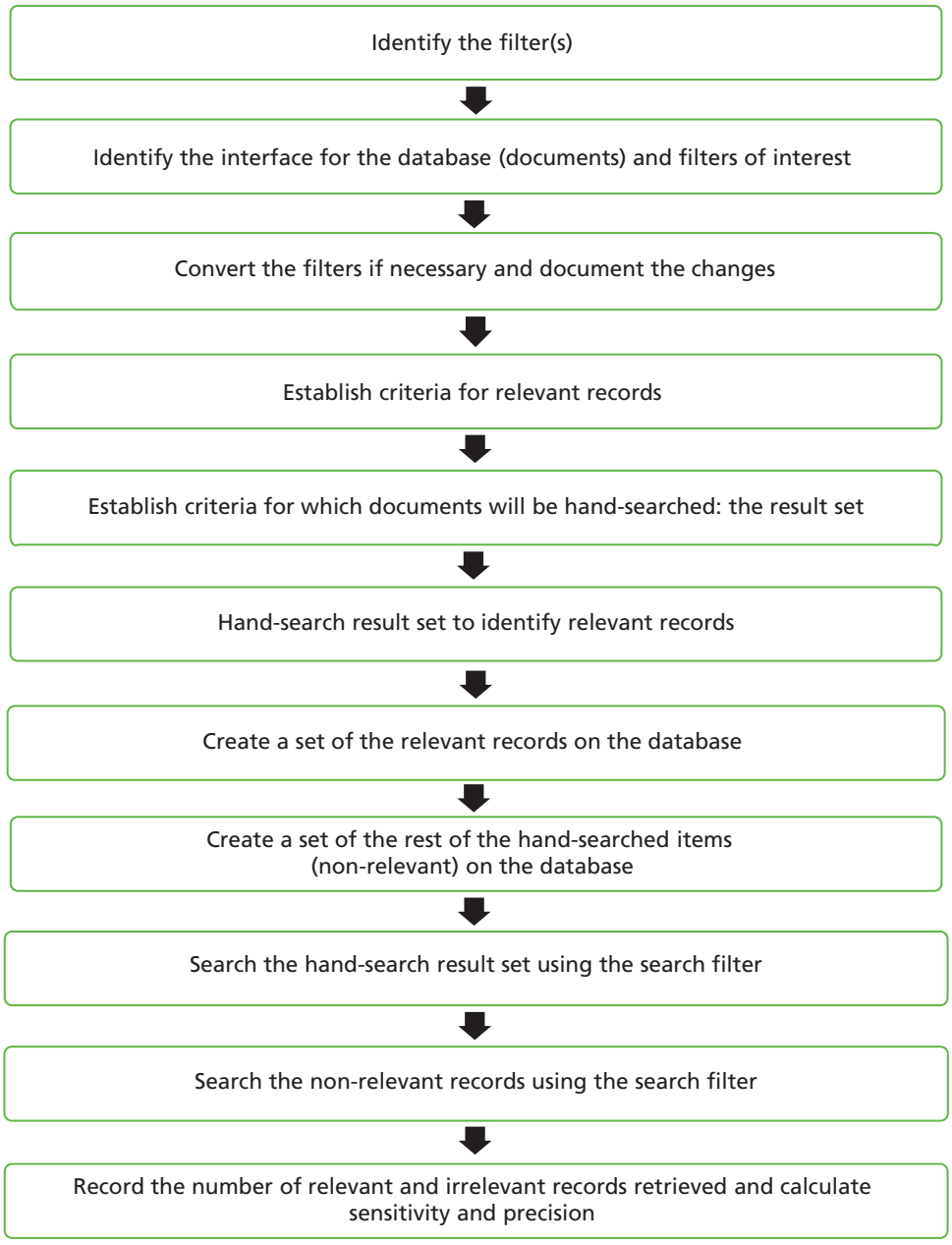

FIGURE 11 Search filter performance measurement using a hand-searched reference set.

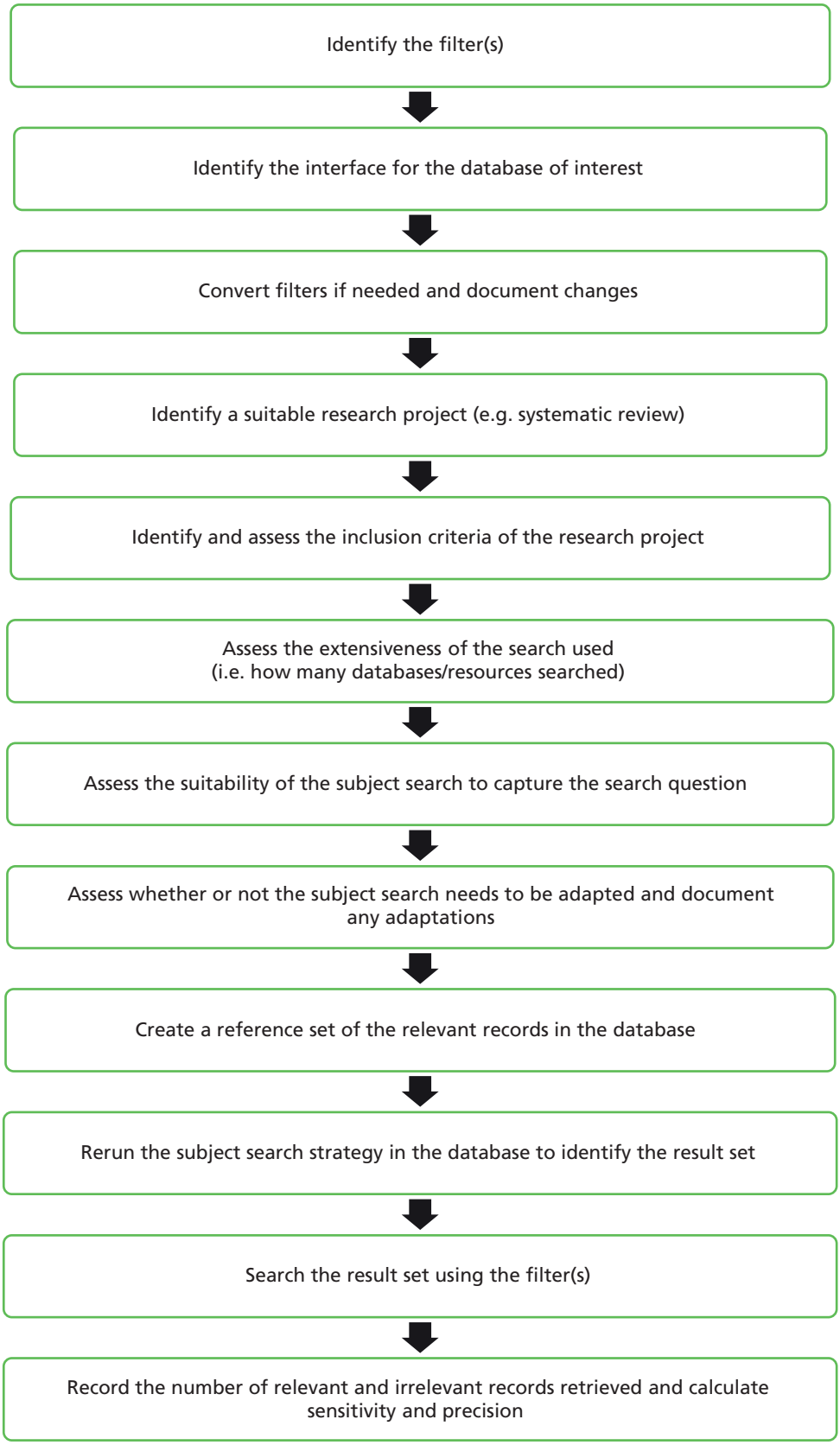

FIGURE 12 Search filter performance measurement using a RR reference set.

Reporting search filter performance

Search filter performance can be reported to the ISSG Search Filters Resource website (see https://sites.google.com/a/york.ac.uk/issg-search-filters-resource/) by e-mailing completed details to the website editors, who will make the data available on the website. A pro forma is provided in *Table 38*, which captures the key data required.

Table 39 provides an example of a completed pro forma.

TABLE 38 Pro forma for reporting search filter performance data

Data element	Details
Filter reference	Bibliographic citation or URL
Filter listing	List the complete filter with syntax here
Database	Database name (e.g. MEDLINE)
Interface (review B indicates that interface is reported only sporadically)	For example Ovid
Comments on the filter	Please indicate any concerns about the filter or any adaptations made to the original
Reference set creation	Hand-search or RR? (see *Hand-searching* and *Relative recall*)
	Describe methods and any limitations
RR: subject search	List subject search here if applicable (see *Relative recall*)
RR: comments on subject search	Please indicate any concerns about the subject search or any adaptations made to the original subject search (see *Relative recall*)
Number of reference set records	Number of relevant records (see *Creating the reference set for testing*)
Number of results records yielded by subject search or hand-search	Number of records returned by the subject search or the total number of records that were hand-searched (see *The results set*)
Number of reference set records retrieved by the filter plus subject search (if subject search is used): sensitivity	Sensitivity of search filter in terms of relevant records (see *Which performance characteristics should be measured?*)
Precision of search filter	Precision (number of reference set records retrieved/number of records in the results set) (see *Which performance characteristics should be measured?*)
Reduction in number needed to screen	See *Measuring search filter performance*
Date of performance test	
Any other comments	

TABLE 39 Example of a completed pro forma

Data element	Details
Filter reference	SIGN DTA filter www.sign.ac.uk/methodology/filters.html#diag (accessed July 2016)
Filter listing	exp "Sensitivity and Specificity"/ sensitivity.tw. specificity.tw. ((pre-test or pretest) adj probability).tw. post-test probability.tw. predictive value$.tw. likelihood ratio$.tw. or/1–7
Database	MEDLINE
Interface (review B indicates that interface is reported only sporadically)	Ovid
Comments on the filter	This filter was used exactly as listed on the SIGN website
Reference set creation	RR. We used the included studies from the HTA review of diagnostic test methods for urinary tract infections:[142] This review was prepared by searching a wide range of resources and using a sensitive search strategy without the use of DTA filters
RR: subject search	1. exp urinary tract infections/ (27,032) 2. bacterial infections/ or exp pseudomonas infections/or exp klebsiella infections/ or gram negative infections/or exp escherichia coli/or exp proteus/ or exp enterococcus/ (217,644) 3. exp staphylococcus/ (41,409) 4. exp leurocytes/ (398,776) 5. (microbial infection? or bacterial infection?).ti,ab. (11,874) 6. (urinary or urine or urethra or bladder or ureter? or kidney or kidneys or renal).ti,ab. (553,654) 7. exp urinary tract/ (251,201) 8. or/2–5 (645,127) 9. or/6–7 (633,796) 10. 8 and 9 (27,291) 11. 1 or 10 (49,809) 12. exp child, preschool/ or exp infant/ (827,649) 13. (infant? or baby or babies or toddler? or preschooler?).ti,ab. (175,142) 14. or/12–13 (857,927) 15. 11 and 14 (7594) 16. (risk assessment? or exam or examination or feeding or slow weight gain or fever or vomiting or diarrh?).ti,ab. (390,002) 17. (((sepsis or failure) adj2 thrive) or malaise or frequent urination or abdominal discomfort or abdominal pain).ti,ab. (20,335) 18. (delayed bladder control or dysuria or (pain adj3 urination) or painful urination or difficult urination).ti,ab. (1587) 19. (urinalysis or urine analysis or urine sample? or urine specimen? or (urine adj3 collect?)).ti,ab. (17,696) 20. (urine bags or dipstick? or dip stick? or urine microscopy).ti,ab. (1074) 21. (reagent strip? or colorimetric test? or gas analysis or impedance or luminescence).ti,ab. (16,858) 22. (immunological test? or elisa or enzyme test? or bacterial oxygen consumption or turbidimetry or urine culture).ti,ab. (48,330) 23. (bacterial culture or dipslide? or renal ultrasonography or planar imaging or radiography or urography or pyelography or kub or bladder imaging).ti,ab. (25,490) 24. (cystography or cystourethrography or nuclear medicine or scintigraphy or cystogram?).ti,ab. (28,553)

continued

TABLE 39 Example of a completed pro forma (*continued*)

Data element	Details
	25. exp physical examination/ or exp fever/ or exp body weight changes/ or exp abdominal pain/ or exp urological manifestations or failure to thrive/ (369,317) 26. exp vomiting/ or diarrhea/ or exp sepsis/ or urinalysis/ (88,329) 27. exp microscopy/ or exp "indicators and reagents"/ (477,710) 28. colorimetry/ or electric impedance/ or exp immunoassay/ or exp fluorescent antibody technique/ (320,665) 29. exp diagnostic imaging/ (811,099) 30. exp nuclear medicine/ or exp cystoscopy/ or exp diagnostic techniques, urological/ (68,980) 31. or/16–30 (2,116,054) 32. 15 and 31 (2893) 33. vesico-ureteral reflux/ or pyelonephritis/ or bacteriuria/ or cystitis/(23,125) 34. (failure adj2 thrive).ti,ab. (2130) 35. sepsis.tw. (28,242) 36. ultrasonography.ti,ab. (30,181) 37. exp succimer/ or exp organometallic compounds/ or technetium/ or exp sulfhydryl compounds/ or exp culture media/ (204,118) 38. urinary catheterization/ or ammonium chloride/ or c-reactive protein/ or urodynamics/ or urine/mi (30758) 39. (dmsa or urogram? or ultrasound? or (renal adj scan?)).ti,ab. (63,607) 40. (spect or (planar adj image?) or (dip adj slide?) or cystoscopy).ti,ab. (12,053) 41. ((bladder adj aspiration) or (acidification adj test?) or (cortical adj echogenicity)).ti,ab. (149) 42. workup.ti,ab. (3809) 43. (radiographic or cystomanometry).ti,ab. (38,227) 44. (bladder adj3 (investigat? or detect?)).ti,ab. (246) 45. (kidney adj3 (investigat? or detect?)).ti,ab. (242) 46. (urethra adj3 (investigat? or detect?)).ti,ab. (7) 47. (renal adj3 (investigat? or detect?)).ti,ab. (984) 48. (kidneys adj3 (investigat? or detect?)).ti,ab. (63) 49. (urinary adj3 (investigat? or detect?)).ti,ab. (479) 50. (infection? adj3 (urinary or urine or urethra or bladder or ureter? or kidney or kidneys or renal)).ti,ab. (22,555) 51. (2 or 3 or 4 or 33) and 7 (14,093) 52. 1 or 50 or 51 (48,512) 53. 52 and 14 (9186) 54. or/34–49 (398,248) 55. 53 and 54 (1988) 56. 55 not 32 (1121)
RR: comments on subject search	This search was used exactly as reported in the review publication
Number of reference set records	187
Number of results records yielded by subject search	1121
Number of reference set records retrieved by the filter plus subject search	150
Number of results records retrieved by the filter plus subject search	1000
Sensitivity (number of reference set records retrieved/number of reference set records)	150/187 = 0.80 or 80%
Precision (number of reference set records retrieved/number of records in the results set)	150/1000 = 0.15 or 15%
Reduction in number needed to screen	1121 – 1000 = 121 fewer records retrieved (reduction of 10.8%)
Date of performance test	28 January 2016
Any other comments	No human restrictions or language restrictions were applied

Chapter 6 Project website

A pilot website relating to the project is available for public access (see https://sites.google.com/a/york.ac.uk/search-filter-performance/).

The website contains extracts from this report (under the headings Abstract, Scientific summary, Aims and objectives, Definitions, Abbreviations and acronyms, Presentations, Publications and Bibliography), together with a link to this report.

The website also contains a test site offering different graphical representations of search filter performance, such as sensitivity, precision and NNR. These representations are in the form of bar charts, scatter plots and radar diagrams.

The website also links to the ISSG Search Filters Resource (see https://sites.google.com/a/york.ac.uk/issg-search-filters-resource/).

Chapter 7 Future research

The following issues have emerged as topics for future research.

Filters for other study designs

- The development and validation of filters for a wider range of study designs, such as epidemiology, quality of life and prognostic studies (questionnaire).
- A review of the performance measures reported for methodological filter performance and performance comparisons for study designs not included in this review would shed light on topics beyond those that we assessed (reviews A and B).

Displaying performance results

- Studies to explore alternative methods of displaying performance results for multiple methodological search filters (reviews A, B and C) and testing of searchers' understanding of the filter performance trade-offs offered (interviews and questionnaire and review E).

Filter amendments

- Translations of filters for different databases and interfaces and the development of more strategies that are independent of indexing language (to facilitate transferral across databases) (questionnaire).
- Qualitative research into exactly how search filters are amended in practice, to inform filter design. Filter designers tend to assume that searchers want sensitive filters or precise filters but in fact searchers may prefer different options or to be able to choose using a sliding scale of sensitivity depending on the number of records retrieved (questionnaire).

Applicability to the wider community

- Interviews with searchers and researchers from other settings to understand whether the NICE experience is generalisable (interviews).

Synthesis of filter performance

- Exploration of methods for the numerical synthesis of the results of several filter performance comparisons (reviews B and C).

Filter-only performance

- Obtain baseline performance for a search filter by running the filter across an entire database (such as MEDLINE) with no subject terms. This removes one of the potential limiting factors of assessing filters in combination with subject searches and also obtains a measure of the prevalence of the study design in the database (see *Chapter 5*).

Acknowledgements

We are grateful to the following individuals for their assistance with this project:

- the co-authors of a review of DTA search filters, which was unpublished at the time of this study but which has since been published,[1,2] for permission to include search filter performance diagrams from their review
- Anne Eisinga (UK Cochrane Centre) for undertaking searches for records for inclusion in the ISSG Search Filters Resource, including an update search for this project
- Tom Hudson (NICE) for participation in a project meeting
- respondents to the interviews and questionnaires
- Mary Edwards and Danielle Varley (YHEC) for providing administrative assistance
- Dianne Wright (YHEC) for setting up the electronic questionnaire
- the (anonymous) peer reviewers for their insightful comments.

We acknowledge that there has been a regrettable delay between carrying out the project, including the searches, and the publication of this report, because of serious illness of the principal investigator. The searches were carried out in 2010/11.

Contributions of authors

Carol Lefebvre (Senior Information Specialist, UK Cochrane Centre) contributed to the drafting of the project proposal, managed the project as the Principal Investigator, co-drafted the report, responded to editors' and peer reviewers' comments and served as guarantor.

Julie Glanville (Information Specialist and Associate Director, YHEC) conceived the project and led in drafting the project proposal, contributed to the management of the project as the Co-Lead Investigator, assisted in the design of the interview schedule and the survey instrument, carried out some of the interviews, co-drafted the report and served as guarantor.

Sophie Beale (Senior Consultant, YHEC) designed the interview schedule and the survey instrument, carried out some of the interviews, analysed the interview and survey data and drafted the interview and survey sections of the report.

Charles Boachie (Statistician, University of Aberdeen) conducted review C and drafted the relevant section of the report.

Steven Duffy (Information Specialist and Research Consultant, YHEC) carried out some of the interviews, conducted review D and drafted the relevant section of the report.

Cynthia Fraser (Information Specialist, University of Aberdeen) conducted reviews C and E, drafted the relevant sections of the report and responded to editors' and peer reviewers' comments.

Jenny Harbour (Information Specialist, Healthcare Improvement Scotland) conducted review B and drafted the relevant section of this report.

Rachael McCool (Research Consultant, YHEC) assisted in the design of the interview schedule and the survey instrument and carried out some of the interviews.

Lynne Smith (Information Specialist, Healthcare Improvement Scotland) conducted review A and drafted the relevant section of the report.

All authors commented on drafts of the interview schedule, the survey instrument, the results of the interviews and survey, and the reviews, and approved a prepublication draft of this manuscript.

Publications

Beale S, Duffy S, Glanville J, Lefebvre C, Wright D, McCool R, *et al.* Choosing and using methodological search filters: searchers' views. *Health Info Libr J* 2014;**31**:133–47.

Harbour J, Fraser C, Lefebvre C, Glanville J, Beale S, Boachie C, *et al.* Reporting methodological search filter performance comparisons: a literature review. *Health Info Libr J* 2014;**31**:176–94.

Presentations

Julie Glanville (YHEC) presented a summary of the project, on behalf of the project team, at the NICE Joint Information Day on 14 November 2011 in London, entitled *MRC-Funded Research Project on Search Filter Performance*.

Jenny Harbour (Healthcare Improvement Scotland) presented aspects of the project, on behalf of the project team, at the LIS DREaM – Developing Research Excellence and Methods Workshop on 30 January 2012 in London. The workshop presentation is available online [see http://lisresearch.org/dream-event-3-unconference-half-hour/ (accessed 28 August 2017)].

Carol Lefebvre (UK Cochrane Centre) presented results from this project, on behalf of the project team, at the HTAi Annual Meeting in Bilbao in June 2012. The poster presentation was entitled *Methodological Search Filters Performance Project: What to Measure and How to Present These Measures?* [poster 249; see www.htai.org/fileadmin/HTAi_Files/Conferences/2012/2012_HTAi_Bilbao_Poster_Presentations.pdf (accessed July 2016)].

Jenny Harbour (Healthcare Improvement Scotland) presented results from this project, on behalf of the project team, at the Health Libraries Group conference in Glasgow in July 2012. Her presentation was entitled 'Search filters performance project: what to measure and how to present these measures?'.

Data sharing statement

All available data and information have been included within this report or added as appendices. Further information can be obtained by contacting the corresponding author.

References

1. 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 Syst Rev* 2013;**9**:MR000022. https://doi.org/10.1002/14651858.MR000022.pub3
2. 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. *J Clin Epidemiol* 2011;**64**:602–7. https://doi.org/10.1016/j.jclinepi.2010.07.006
3. Bak G, Mierzwinski-Urban M, Fitzsimmons H, Morrison A, Maden-Jenkins M. A pragmatic critical appraisal instrument for search filters: introducing the CADTH CAI. *Health Info Libr J* 2009;**26**:211–19. https://doi.org/10.1111/j.1471-1842.2008.00830.x
4. Glanville J, Bayliss S, Booth A, Dundar Y, Fernandes H, Fleeman ND, *et al.* So many filters, so little time: the development of a search filter appraisal checklist. *J Med Libr Assoc* 2008;**96**:356–61. https://doi.org/10.3163/1536-5050.96.4.011
5. Jenkins M, Johnson F. Awareness, use and opinions of methodological search filters used for the retrieval of evidence-based medical literature – a questionnaire survey. *Health Info Libr J* 2004;**21**:33–43. https://doi.org/10.1111/j.1471-1842.2004.00480.x
6. Glanville J, Lefebvre C, Wright K. *ISSG Search Filter Resource*. York: The InterTASC Information Specialists' Sub-Group; 2008 [updated 2017]. URL: https://sites.google.com/a/york.ac.uk/issg-search-filters-resource/home (accessed 22 August 2017).
7. McKinlay RJ, Wilczynski NL, Haynes RB, Hedges team. Optimal search strategies for detecting cost and economic studies in EMBASE. *BMC Health Serv Res* 2006;**6**:67. https://doi.org/10.1186/1472-6963-6-67
8. Wilczynski NL, Haynes RB, Lavis JN, Ramkissoonsingh R, Arnold-Oatley AE, HSR Hedges team. Optimal search strategies for detecting health services research studies in MEDLINE. *CMAJ* 2004;**171**:1179–85. https://doi.org/10.1503/cmaj.1040512
9. 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**:365–73. https://doi.org/10.1148/radiol.2472070101
10. Bachmann LM, Coray R, Estermann P, Ter Riet G. Identifying diagnostic studies in MEDLINE: reducing the number needed to read. *J Am Med Inform Assoc* 2002;**9**:653–8. https://doi.org/10.1197/jamia.M1124
11. Bachmann LM, Estermann P, Kronenberg C, ter Riet G. Identifying diagnostic accuracy studies in EMBASE. *J Med Libr Assoc* 2003;**91**:341–6.
12. 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. *Int J Nurs Terminol Classif* 2005;**16**:26–32. https://doi.org/10.1111/j.1744-618X.2005.00006.x
13. Haynes RB, McKibbon KA, Wilczynski NL, Walter SD, Werre SR. Optimal search strategies for retrieving scientifically strong studies of treatment from MEDLINE: analytical survey. *BMJ* 2004;**328**:1040. https://doi.org/10.1136/bmj.38068.557998.EE
14. Vincent S, Greenley S, Beaven O. Clinical evidence diagnosis: developing a sensitive search strategy to retrieve diagnostic studies on deep vein thrombosis: a pragmatic approach. *Health Info Libr J* 2003;**20**:150–9. https://doi.org/10.1046/j.1365-2532.2003.00427.x

15. Wilczynski NL, Haynes RB, Hedges team. EMBASE search strategies for identifying methodologically sound diagnostic studies for use by clinicians and researchers. *BMC Med Res Methodol* 2005;**3**:7. https://doi.org/10.1186/1741-7015-3-7

16. Eady AM, Wilczynski NL, Haynes RB. PsycINFO search strategies identified methodologically sound therapy studies and review articles for use by clinicians and researchers. *J Clin Epidemiol* 2008;**61**:34–40. https://doi.org/10.1016/j.jclinepi.2006.09.016

17. Montori VM, Wilczynski NL, Morgan D, Haynes RB, Hedges team. Optimal search strategies for retrieving systematic reviews from MEDLINE: analytical survey. *BMJ* 2005;**330**:68. https://doi.org/10.1136/bmj.38336.804167.47

18. Shojania KG, Bero LA. Taking advantage of the explosion of systematic reviews: an efficient MEDLINE search strategy. *Eff Clin Pract* 2001;**4**:157–62.

19. White VJ, Glanville JM, Klefebvre C, Sheldon TA. A statistical approach to designing search filters to find systematic reviews: objectivity enhances accuracy. *J Info Sci* 2001;**27**:357–70. https://doi.org/10.1177/016555150102700601

20. Wilczynski NL, Haynes RB, Hedges team. EMBASE search strategies achieved high sensitivity and specificity for retrieving methodologically sound systematic reviews. *J Clin Epidemiol* 2007;**60**:29–33. https://doi.org/10.1016/j.jclinepi.2006.04.001

21. Wong SS, Wilczynski NL, Haynes RB. Optimal CINAHL search strategies for identifying therapy studies and review articles. *J Nurs Scholarsh* 2006;**38**:194–9. https://doi.org/10.1111/j.1547-5069.2006.00100.x

22. Glanville JM, Lefebvre C, Miles JN, Camosso-Stefinovic J. How to identify randomized controlled trials in MEDLINE: ten years on. *J Med Libr Assoc* 2006;**94**:130–6.

23. 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**:21. https://doi.org/10.1136/bmj.38446.498542.8F

24. Lefebvre C, Eisinga A, McDonald S, Paul N. Enhancing access to reports of randomized trials published world-wide – the contribution of EMBASE records to the Cochrane Central Register of Controlled Trials (CENTRAL) in The Cochrane Library. *Emerg Themes Epidemiol* 2008;**5**:13. https://doi.org/10.1186/1742-7622-5-13

25. Manríquez JJ. A highly sensitive search strategy for clinical trials in Literatura Latino Americana e do Caribe em Ciências da Saúde (LILACS) was developed. *J Clin Epidemiol* 2008;**61**:407–11. https://doi.org/10.1016/j.jclinepi.2007.06.009

26. Robinson KA, Dickersin K. Development of a highly sensitive search strategy for the retrieval of reports of controlled trials using PubMed. *Int J Epidemiol* 2002;**31**:150–3. https://doi.org/10.1093/ije/31.1.150

27. Taljaard M, McGowan J, Grimshaw JM, Brehaut JC, McRae A, Eccles MP, Donner A. Electronic search strategies to identify reports of cluster randomized trials in MEDLINE: low precision will improve with adherence to reporting standards. *BMC Med Res Methodol* 2010;**10**:15. https://doi.org/10.1186/1471-2288-10-15

28. Wong SS, Wilczynski NL, Haynes RB. Developing optimal search strategies for detecting clinically sound treatment studies in EMBASE. *J Med Libr Assoc* 2006;**94**:41–7.

29. Zhang L, Ajiferuke I, Sampson M. Optimizing search strategies to identify randomized controlled trials in MEDLINE. *BMC Med Res Methodol* 2006;**6**:23. https://doi.org/10.1186/1471-2288-6-23

30. Abhijnhan A, Surcheva Z, Wright J, Adams CE. Searching a biomedical bibliographic database from Bulgaria: the ABS database. *Health Info Libr J* 2007;**24**:200–3. https://doi.org/10.1111/j.1471-1842.2007.00723.x

31. Almerie MQ, Matar HE, Jones V, Kumar A, Wright J, Wlostowska E, Adams CE. Searching the Polish Medical Bibliography (Polska Bibliografia Lekarska) for trials. *Health Info Libr J* 2007;**24**:283–6. https://doi.org/10.1111/j.1471-1842.2007.00716.x

32. Chow TK, To E, Goodchild CS, McNeil JJ. A simple, fast, easy method to identify the evidence base in pain-relief research: validation of a computer search strategy used alone to identify quality randomized controlled trials. *Anesth Analg* 2004;**98**:1557–65. https://doi.org/10.1213/01.ANE.0000114071.78448.2D

33. Corrao S, Colomba D, Arnone S, Argano C, Di Chiara T, Scaglione R, Licata G. Improving efficacy of PubMed clinical queries for retrieving scientifically strong studies on treatment. *J Am Med Inform Assoc* 2006;**13**:485–7. https://doi.org/10.1197/jamia.M2084

34. Day D, Furlan A, Irvin E, Bombardier C. Simplified search strategies were effective in identifying clinical trials of pharmaceuticals and physical modalities. *J Clin Epidemiol* 2005;**58**:874–81. https://doi.org/10.1016/j.jclinepi.2005.02.005

35. de Freitas AE, Herbert RD, Latimer J, Ferreira PH. Searching the LILACS database for Portuguese- and Spanish-language randomized trials in physiotherapy was difficult. *J Clin Epidemiol* 2005;**58**:233–7. https://doi.org/10.1016/j.jclinepi.2004.06.014

36. Devillé WL, Buntinx F, Bouter LM, Montori VM, de Vet HCW, van der Windt D, Bezemer PD. Conducting systematic reviews of diagnostic studies: didactic guidelines. *BMC Med Res Methodol* 2002;**2**:9. https://doi.org/10.1186/1471-2288-2-9

37. Eisinga A, Siegfried N, Clarke M. The sensitivity and precision of search terms in Phases I, II and III of the Cochrane Highly Sensitive Search Strategy for identifying reports of randomized trials in MEDLINE in a specific area of health care – HIV/AIDS prevention and treatment interventions. *Health Info Libr J* 2007;**24**:103–9. https://doi.org/10.1111/j.1471-1842.2007.00698.x

38. Kele I, Bereczki D, Furtado V, Wright J, Adams CE. Searching a biomedical bibliographic database from Hungary – the 'Magyar Orvosi Bibliografia'. *Health Info Libr J* 2005;**22**:286–95. https://doi.org/10.1111/j.1471-1842.2005.00577.x

39. Kumar A, Wright J, Adams CE. Searching a biomedical bibliographic database from the Ukraine: the Panteleimon database. *Health Info Libr J* 2005;**22**:223–7. https://doi.org/10.1111/j.1471-1842.2005.00578.x

40. McDonald S. Improving access to the international coverage of reports of controlled trials in electronic databases: a search of the Australasian Medical Index. *Health Info Libr J* 2002;**19**:14–20. https://doi.org/10.1046/j.0265-6647.2001.00359.x

41. Royle P, Waugh N. Literature searching for clinical and cost-effectiveness studies used in health technology assessment reports carried out for the National Institute for Clinical Excellence appraisal system. *Health Technology Assessment* 2003;**7**(34). https://doi.org/10.3310/hta7340

42. Royle P, Waugh N. A simplified search strategy for identifying randomised controlled trials for systematic reviews of health care interventions: a comparison with more exhaustive strategies. *BMC Med Res Methodol* 2005;**5**:23. https://doi.org/10.1186/1471-2288-5-23

43. Royle P, Waugh N. Making literature searches easier: a rapid and sensitive search filter for retrieving randomized controlled trials from PubMed. *Diabet Med* 2007;**24**:308–11. https://doi.org/10.1111/j.1464-5491.2007.02046.x

44. Sassi F, Archard L, McDaid D. Searching literature databases for health care economic evaluations: how systematic can we afford to be? *Med Care* 2002;**40**:387–94. https://doi.org/10.1097/00005650-200205000-00004

45. Wilczynski NL, Haynes RB. Consistency and accuracy of indexing systematic review articles and meta-analyses in Medline. *Health Info Libr J* 2009;**26**:203–10. https://doi.org/10.1111/j.1471-1842.2008.00823.x

46. Harbour J, Fraser C, Lefebvre C, Glanville J, Beale S, Boachie C, *et al.* Reporting methodological search filter performance comparisons: a literature review. *Health Info Libr J* 2014;**31**:176–94. https://doi.org/10.1111/hir.12070

47. Glanville J, Fleetwood K, Yellowlees A, Kaunelis D, Mensinkai S. *Development and Testing of Search Filters to Identify Economic Evaluations in MEDLINE and EMBASE*. Ottawa, ON: Canadian Agency for Drugs and Technologies in Health (CADTH); 2009.

48. 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. *J Clin Epidemiol* 2006;**59**:234–40. https://doi.org/10.1016/j.jclinepi.2005.07.014

49. Ritchie G, Glanville J, Lefebvre C. Do published search filters to identify diagnostic test accuracy studies perform adequately? *Health Info Libr J* 2007;**24**:188–92. https://doi.org/10.1111/j.1471-1842.2007.00735.x

50. Deurenberg R, Vlayen J, Guillo S, Oliver TK, Fervers B, Burgers J. Standardization of search methods for guideline development: an international survey of evidence-based guideline development groups. *Health Info Libr J* 2008;**25**:23–30. https://doi.org/10.1111/j.1471-1842.2007.00732.x

51. Jenkins M. Evaluation of methodological search filters – a review. *Health Info Libr J* 2004;**21**:148–63. https://doi.org/10.1111/j.1471-1842.2004.00511.x

52. Sampson M, Zhang L, Morrison A, Barrowman NJ, Clifford TJ, Platt RW, *et al.* An alternative to the hand searching gold standard: validating methodological search filters using relative recall. *BMC Med Res Methodol* 2006;**6**:33. https://doi.org/10.1186/1471-2288-6-33

53. Boluyt N, Tkosvold L, Lefebvre C, Klassen TP, Offringa M. The usefulness of systematic review search strategies in finding child health systematic reviews in MEDLINE. *Arch Pediatr Adolesc Med* 2008;**162**:111–16. https://doi.org/10.1001/archpediatrics.2007.40

54. Royle P, Milne R. Literature searching for randomized controlled trials used in Cochrane reviews: rapid versus exhaustive searches. *Int J Technol Assess Health Care* 2003;**19**:591–603. https://doi.org/10.1017/S0266462303000552

55. Bardia A, Wahner-Roedler DL, Erwin PL, Sood A. Search strategies for retrieving complementary and alternative medicine clinical trials in oncology. *Integr Cancer Ther* 2006;**5**:202–5. https://doi.org/10.1177/1534735406292146

56. Boynton J, Glanville J, McDaid D, Lefebvre C. Identifying systematic reviews in MEDLINE: developing an objective approach to search strategy design. *J Info Sci* 1998;**24**:137–57. https://doi.org/10.1177/016555159802400301

57. Devillé WL, Bezemer PD, Bouter LM. Publications on diagnostic test evaluation in family medicine journals: an optimal search strategy. *J Clin Epidemiol* 2000;**53**:65–9. https://doi.org/10.1016/S0895-4356(99)00144-4

58. 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. *J Clin Epidemiol* 2005;**58**:444–9. https://doi.org/10.1016/j.jclinepi.2004.09.011

59. Glanville J, Kaunelis D, Mensinkai S. How well do search filters perform in identifying economic evaluations in MEDLINE and EMBASE. *Int J Technol Assess Health Care* 2009;**25**:522–9. https://doi.org/10.1017/S0266462309990523

60. 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. *J Clin Epidemiol* 2009;**62**:974–81. https://doi.org/10.1016/j.jclinepi.2008.11.006

61. McKibbon KA, Wilczynski NL, Haynes RB, Hedges team. Retrieving randomized controlled trials from MEDLINE: a comparison of 38 published search filters. *Health Info Libr J* 2009;**26**:187–202. https://doi.org/10.1111/j.1471-1842.2008.00827.x

62. Royle P, Waugh N. A simplified search strategy for identifying randomised controlled trials for systematic reviews of health care interventions: a comparison with more exhaustive strategies. *BMC Med Res Methodol* 2005;**5**:2–3. https://doi.org/10.1186/1471-2288-5-2

63. Wong SS, Wilczynski NL, Haynes RB. Comparison of top-performing search strategies for detecting clinically sound treatment studies and systematic reviews in MEDLINE and EMBASE. *J Med Libr Assoc* 2006;**94**:451–5.

64. Haynes RB, Wilczynski N, McKibbon KA, Walker CJ, Sinclar JC. Developing optimal search strategies for detecting clinically sound studies in MEDLINE. *J Am Med Inform Assoc* 1994;**1**:447–58.

65. Castro AA, Clark OA, Atallah AN. Optimal search strategy for clinical trials in the Latin American and Caribbean Health Science Literature database (LILACS database): update. *Sao Paulo Med J* 1999;**117**:138e9. http://dx.doi.org/10.1590/S1516-31801997000300004

66. Bradley SM. Examination of the clinical queries and systematic review 'hedges' in EMBASE and MEDLINE. *J Can Health Libr Assoc* 2010;**31**:27–37. https://doi.org/10.5596/c10-022

67. Leeflang MM, Deeks JJ, Gatsonis C, Bossuyt PM, Cochrane Diagnostic Test Accuracy Working Group. Systematic reviews of diagnostic test accuracy. *Ann Intern Med* 2008;**149**:889–97. https://doi.org/10.7326/0003-4819-149-12-200812160-00008

68. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, *et al.* Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. *BMJ* 2003;**326**:41–4. https://doi.org/10.1136/bmj.326.7379.41

69. Centre for Reviews and Dissemination. *Systematic Reviews: CRD's Guidance for Undertaking Reviews in Health Care*. York: University of York; 2009.

70. Hui SL, Zhou XH. Evaluation of diagnostic tests without gold standards. *Stat Methods Med Res* 1998;**7**:354–70. https://doi.org/10.1177/096228029800700404

71. Reitsma JB, Rutjes AW, Khan KS, Coomarasamy A, Bossuyt PM. A review of solutions for diagnostic accuracy studies with an imperfect or missing reference standard. *J Clin Epidemiol* 2009;**62**:797–806. https://doi.org/10.1016/j.jclinepi.2009.02.005

72. Rutjes AW, Reitsma JB, Coomarasamy A, Khan KS, Bossuyt PM. Evaluation of diagnostic tests when there is no gold standard. A review of methods. *Health Technol Assess* 2007;**11**(50). https://doi.org/10.3310/hta11500

73. Bossuyt P, Leeflang M. Chapter 6: developing criteria for including studies. In *Cochrane Handbook of Systematic Reviews of Diagnostic Test Accuracy Version 0.4*. The Cochrane Collaboration; 2008.

74. Deeks JJ. Systematic reviews in health care: systematic reviews of evaluations of diagnostic and screening tests. *BMJ* 2001;**323**:157–62. https://doi.org/10.1136/bmj.323.7305.157

75. Food and Drug Administration. *Guidance for Industry and Staff: Statistical Guidance on Reporting Results from Studies Evaluating Diagnostic Tests*. US Department of Health and Human Services; 2007. URL: www.fda.gov/default.htm (accessed 1 June 2011).

76. Medical Services Advisory Committee. *Guidelines for the Assessment of Diagnostic Technologies*. Canberra, ACT: Australian Government Department of Health and Ageing; 2005.

77. Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, van der Meulen JH, Bossuyt PM. Empirical evidence of design-related bias in studies of diagnostic tests. *JAMA* 1999;**282**:1061–6. https://doi.org/10.1001/jama.282.11.1061

78. Westwood ME, Whiting PF, Kleijnen J. How does study quality affect the results of a diagnostic meta-analysis? *BMC Med Res Methodol* 2005;**5**:20. https://doi.org/10.1186/1471-2288-5-20

79. Whiting P, Rutjes AW, Reitsma JB, Glas AS, Bossuyt PM, Kleijnen J. Sources of variation and bias in studies of diagnostic accuracy: a systematic review. *Ann Intern Med* 2004;**140**:189–202. https://doi.org/10.7326/0003-4819-140-3-200402030-00010

80. Cook C, Cleland J, Huijbregts P. Creation and critique of studies of diagnostic accuracy: use of the STARD and QUADAS methodological quality assessment tools. *J Man Manip Ther* 2007;**15**:93–102. https://doi.org/10.1179/106698107790819945

81. Bachmann LM, Puhan MA, ter Riet G, Bossuyt PM. Sample sizes of studies on diagnostic accuracy: literature survey. *BMJ* 2006;**332**:1127–9. https://doi.org/10.1136/bmj.38793.637789.2F

82. Bochmann F, Johnson Z, Azuara-Blanco A. Sample size in studies on diagnostic accuracy in ophthalmology: a literature survey. *Br J Ophthalmol* 2007;**91**:898–900. https://doi.org/10.1136/bjo.2006.113290

83. Flahault A, Cadilhac M, Thomas G. Sample size calculation should be performed for design accuracy in diagnostic test studies. *J Clin Epidemiol* 2005;**58**:859–62. https://doi.org/10.1016/j.jclinepi.2004.12.009

84. Whiting P, Rutjes AW, Dinnes J, Reitsma J, Bossuyt PM, Kleijnen J. Development and validation of methods for assessing the quality of diagnostic accuracy studies. *Health Technol Assess* 2004;**8**(25). https://doi.org/10.3310/hta8250

85. Whiting P, Rutjes AW, Reitsma JB, Bossuyt PM, Kleijnen J. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. *BMC Med Res Methodol* 2003;**3**:25. https://doi.org/10.1186/1471-2288-3-25

86. Belgian Health Care Knowledge Centre. *HTA: Molecular Diagnostics in Belgium*. Report no. 20A. Brussels: Federaal Kenniscentrum voor de Gezondheidszorg; 2005. URL: www.kce.fgov.be/ (accessed 1 June 2011).

87. Whiting PF, Sterne JA, Westwood ME, Bachmann LM, Harbord R, Egger M, Deeks JJ. Graphical presentation of diagnostic information. *BMC Med Res Methodol* 2008;**8**:20. https://doi.org/10.1186/1471-2288-8-20

88. Jaeschke R, Guyatt GH, Sackett DL. Users' guides to the medical literature. III. How to use an article about a diagnostic test. B. What are the results and will they help me in caring for my patients? The Evidence-Based Medicine Working Group. *JAMA* 1994;**271**:703–7. https://doi.org/10.1001/jama.1994.03510330081039

89. Deeks J, Altman DG, Bradburn MJ. Chapter 10: Statistical Methods for Examining Heterogeneity and Combining Results from Several Studies in Meta-analysis. In Egger M, Davey Smith G, Altman DG, editors. *Systematic Reviews in Health Care: Meta-analysis in Context*. London: BMJ Publishing Group; 2001. pp. 285–312. https://doi.org/10.1002/9780470693926.ch15

90. Deeks J, Bossuyt P, Gatsonis C, Macaskill P, Harbord R, Takwoingi Y. Analysing and Presenting Results. In *Cochrane Handbook of Systematic Reviews of Diagnostic Test Accuracy* Version 0.9.0. The Cochrane Collaboration; 2010.

91. Bossuyt PM. The quality of reporting in diagnostic test research: getting better, still not optimal. *Clin Chem* 2004;**50**:465–6. https://doi.org/10.1373/clinchem.2003.029736

92. Coppus SF, van der Veen F, Bossuyt PM, Mol BW. Quality of reporting of test accuracy studies in reproductive medicine: impact of the Standards for Reporting of Diagnostic Accuracy (STARD) initiative. *Fertil Steril* 2006;**86**:1321–9. https://doi.org/10.1016/j.fertnstert.2006.03.050

93. Harper R, Reeves B. Reporting of precision of estimates for diagnostic accuracy: a review. *BMJ* 1999;**318**:1322–3. https://doi.org/10.1136/bmj.318.7194.1322

94. Rama KR, Poovali S, Apsingi S. Quality of reporting of orthopaedic diagnostic accuracy studies is suboptimal. *Clin Orthop Relat Res* 2006;**447**:237–46. https://doi.org/10.1097/01.blo.0000205906.44103.a3

95. Shunmugam M, Azuara-Blanco A. The quality of reporting of diagnostic accuracy studies in glaucoma using the Heidelberg retina tomograph. *Invest Ophthalmol Vis Sci* 2006;**47**:2317–23. https://doi.org/10.1167/iovs.05-1250

96. Siddiqui MA, Azuara-Blanco A, Burr J. The quality of reporting of diagnostic accuracy studies published in ophthalmic journals. *Br J Ophthalmol* 2005;**89**:261–5. https://doi.org/10.1136/bjo.2004.051862

97. Smidt N, Rutjes AW, van der Windt DA, Ostelo RW, Bossuyt PM, Reitsma JB. Reproducibility of the STARD checklist: an instrument to assess the quality of reporting of diagnostic accuracy studies. *BMC Med Res Method* 2006;**6**:12. https://doi.org/10.1186/1471-2288-6-12

98. Wilczynski NL. Quality of reporting of diagnostic accuracy studies: no change since STARD statement publication – before-and-after study. *Radiology* 2008;**248**:817–23. https://doi.org/10.1148/radiol.2483072067

99. Smidt N, Rutjes AW, van der Windt DA, Ostelo RW, Bossuyt PM, Reitsma JB, *et al.* The quality of diagnostic accuracy studies since the STARD statement: has it improved? *Neurology* 2006;**67**:792–7. https://doi.org/10.1212/01.wnl.0000238386.41398.30

100. Agency for Healthcare Research and Quality. A *Comprehensive Overview of the Methods and Reporting of Meta-analyses of Test Accuracy*. Rockville, MD: Agency for Healthcare Research and Quality; 2011.

101. Honest H, Khan KS. Reporting of measures of accuracy in systematic reviews of diagnostic literature. *BMC Health Serv Res* 2002;**2**:4. https://doi.org/10.1186/1472-6963-2-4

102. Mallett S, Deeks JJ, Halligan S, Hopewell S, Cornelius V, Altman DG. Systematic reviews of diagnostic tests in cancer: review of methods and reporting. *BMJ* 2006;**333**:413. https://doi.org/10.1136/bmj.38895.467130.55

103. Belgian Health Care Knowledge C. *Search for Evidence and Critical Appraisal: Health Technology Assessment* [*Process Notes* D200/710.*273/40*]. Brussels: Belgian Health Care Knowledge Centre; 2007.

104. National Institute for Health and Care Excellence. *Interim Methods Statement: Centre for Health Technology Evaluation, Diagnostics Assessment Programme*. London: NICE; 2010. URL: www.nice.org.uk/ (accessed 1 June 2011).

105. Morris RK, Selman TJ, Zamora J, Khan KS. Methodological quality of test accuracy studies included in systematic reviews in obstetrics and gynaecology: sources of bias. *BMC Womens Health* 2011;**11**:7. https://doi.org/10.1186/1472-6874-11-7

106. Whiting P, Rutjes AW, Dinnes J, Reitsma JB, Bossuyt PM, Kleijnen J. A systematic review finds that diagnostic reviews fail to incorporate quality despite available tools. *J Clin Epidemiol* 2005;**58**:1–12. https://doi.org/10.1016/j.jclinepi.2004.04.008

107. Leeflang MM, Bossuyt PM, Irwig L. Diagnostic test accuracy may vary with prevalence: implications for evidence-based diagnosis. *J Clin Epidemiol* 2009;**62**:5–12. https://doi.org/10.1016/j.jclinepi.2008.04.007

108. Rutjes AW, Reitsma JB, Di Nisio M, Smidt N, van Rijn JC, Bossuyt PM. Evidence of bias and variation in diagnostic accuracy studies. *CMAJ* 2006;**174**:469–76. https://doi.org/10.1503/cmaj.050090

109. Dinnes J, Deeks J, Kirby J, Roderick P. A methodological review of how heterogeneity has been examined in systematic reviews of diagnostic test accuracy. *Health Technol Assess* 2005;**9**(12). https://doi.org/10.3310/hta9120

110. Reitsma JB, Glas AS, Rutjes AW, Scholten RJ, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *J Clin Epidemiol* 2005;**58**:982–90. https://doi.org/10.1016/j.jclinepi.2005.02.022

111. Rutter CM, Gatsonis CA. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. *Stat Med* 2001;**20**:2865–84. https://doi.org/10.1002/sim.942

112. Willis BH, Quigley M. Uptake of newer methodological developments and the deployment of meta-analysis in diagnostic test research: a systematic review. *BMC Med Res Methodol* 2011;**11**:27. https://doi.org/10.1186/1471-2288-11-27

113. Davis J, Goadrich M. *The Relationship between Precision-Recall and ROC Curves*. Proceedings of the 23rd Annual Conference on Machine Learning, Pittsburgh, PA, 2006. New York, NY: ACM; 2006. pp. 233–40. https://doi.org/10.1145/1143844.1143874

114. Higgins J, Green S, editors. *Cochrane Handbook for Systematic Reviews of Interventions* Version 5.1.0 [updated March 2011]. The Cochrane Collaboration; 2011. URL: www.handbook.cochrane.org (accessed 25 August 2017).

115. Jha S, Ho A, Bhargavan M, Owen JB, Sunshine JH. Imaging evaluation for suspected pulmonary embolism: what do emergency physicians and radiologists say? *AJR Am J Roentgenol* 2010;**194**:W38–48. https://doi.org/10.2214/AJR.09.2694

116. McGinnis PQ, Hack LM, Nixon-Cave K, Michlovitz SL. Factors that influence the clinical decision making of physical therapists in choosing a balance assessment approach. *Phys Ther* 2009;**89**:233–47. https://doi.org/10.2522/ptj.20080131

117. Perneger TV, Martin DP, Bovier PA. Physicians' attitudes toward health care rationing. *Med Decis Making* 2002;**22**:65–70. https://doi.org/10.1177/0272989X0202200106

118. Sox CM, Koepsell TD, Doctor JN, Christakis DA. Pediatricians' clinical decision making: results of 2 randomized controlled trials of test performance characteristics. *Arch Pediatr Adolesc Med* 2006;**160**:487–92. https://doi.org/10.1001/archpedi.160.5.487

119. Stein PD, Sostman HD, Dalen JE, Bailey DL, Bajc M, Goldhaber SZ, *et al.* Controversies in diagnosis of pulmonary embolism. *Clin Appl Thromb Hemost* 2011;**17**:140–9. https://doi.org/10.1177/1076029610389027

120. Wackerbarth SB, Tarasenko YN, Curtis LA, Joyce JM, Haist SA. Using decision tree models to depict primary care physicians CRC screening decision heuristics. *J Gen Intern Med* 2007;**22**:1467–9. https://doi.org/10.1007/s11606-007-0338-6

121. Zettler M, Mollon B, da Silva V, Howe B, Speechley M, Vinden C. Family physicians' choices of and opinions on colorectal cancer screening modalities. *Can Fam Physician* 2010;**56**:e338–44.

122. UK National Screening Committee. *Criteria for Appraising the Viability, Effectiveness and Appropriateness of a Screening Programme*. 2011. URL: www.screening.nhs.uk/criteria (accessed 1 June 2011).

123. US Preventive Services Task Force. *Procedure Manual*. Publication No. 08-05118-EF. US Preventive Services Task Force; 2008. URL: www.uspreventiveservicestaskforce.org/uspstf08/methods/procmanual.htm (accessed 1 June 2011).

124. Australian Population Health Development Principal Committee Screening Subcommittee. *Population Based Screening Framework*. Canberra, ACT, Australian Population Health Development Principal Committee Screening Subcommittee; 2008. URL: www.health.gov.au (accessed June 2011).

125. World Health Organization. *Screening for Various Cancers*. Geneva: WHO; 2011. URL: www.who.int/cancer/detection/variouscancer/en/ (accessed 1 June 2011).

126. Agoritsas T, Courvoisier DS, Combescure C, Deom M, Perneger TV. Does prevalence matter to physicians in estimating post-test probability of disease? A randomized trial. *J Gen Intern Med* 2011;**26**:373–8. https://doi.org/10.1007/s11606-010-1540-5

127. Bramwell R, West H, Salmon P. Health professionals' and service users' interpretation of screening test results: experimental study. *BMJ* 2006;**333**:284. https://doi.org/10.1136/bmj.38884.663102.AE

128. Cahan A, Gilon D, Manor O, Paltiel O. Probabilistic reasoning and clinical decision-making: do doctors overestimate diagnostic probabilities? *Q J Med* 2003;**96**:763–9. https://doi.org/10.1093/qjmed/hcg122

129. Heller RF, Sandars JE, Patterson L, McElduff P. GPs' and physicians' interpretation of risks, benefits and diagnostic test results. *Fam Pract* 2004;**21**:155–9. https://doi.org/10.1093/fampra/cmh209

130. Sox CM, Doctor JN, Koepsell TD, Christakis DA. The influence of types of decision support on physicians' decision making. *Arch Dis Child* 2009;**94**:185–90. https://doi.org/10.1136/adc.2008.141903

131. Steurer J, Fischer JE, Bachmann LM, Koller M, ter Riet G. Communicating accuracy of tests to general practitioners: a controlled study. *BMJ* 2002;**324**:824–6. https://doi.org/10.1136/bmj.324.7341.824

132. Beale S, Duffy S, Glanville J, Lefebvre C, Wright D, McCool R, *et al.* Choosing and using methodological search filters: searchers' views. *Health Info Libr J* 2014;**31**:133–47. https://doi.org/10.1111/hir.12062

133. 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. URL: http://training.cochrane.org/handbook (accessed 22 September 2017).

134. EBLIP Editorial Team. *Evidence Based Library and Information Practice*. Edmonton, AB: University of Alberta; 2014. URL: http://ejournals.library.ualberta.ca/index.php/EBLIP/index (accessed 1 September 2014).

135. Hausner E, Waffenschmidt S, Kaiser T, Simon M. Routine development of objectively derived search strategies. *Syst Rev* 2012;**1**:19. https://doi.org/10.1186/2046-4053-1-19

136. Noyes P, Popay J, Pearson A, Hannes K, Booth A. Chapter 20: Qualitative Research and Cochrane Reviews. In Higgins J, Green S, editors. *Cochrane Handbook for Systematic Reviews of Interventions* Version 5.1.0 [updated March 2011]. The Cochrane Collaboration; 2011. URL: http://training.cochrane.org/handbook (accessed 22 September 2017).

137. Sampson M, Tetzlaff J, Urquhart C. Precision of healthcare systematic review searches in a cross-sectional sample. *Res Synth Methods* 2011;**2**:119–25. https://doi.org/10.1002/jrsm.42

138. Hopewell S, Clark M, Lefebvre C, Scherer R. Handsearching still a valuable element of the systematic review. *Evid Based Dent* 2008;**9**:85. https://doi.org/10.1038/sj.ebd.6400602

139. Hopewell S, Clarke M, Lefebvre C, Scherer R. Handsearching versus electronic searching to identify reports of randomized trials. *Cochrane Database Syst Rev* 2007;**2**:MR000001. https://doi.org/10.1002/14651858.MR000001.pub2

140. McGowan J, Sampson M, Salzwedel DM, Cogo E, Foerster V, Lefebvre C. PRESS Peer Review of Electronic Search Strategies: 2015 guideline statement. *J Clin Epidemiol* 2016;**75**:40–6. https://doi.org/10.1016/j.jclinepi.2016.01.021

141. McGowan J, Sampson M, Salzwedel DM, Cogo E, Foerster V, Lefebvre C. *PRESS – Peer Review of Electronic Search Strategies: 2015 Guideline Explanation and Elaboration (PRESS E&E)*. Ottawa, ON: Canadian Agency for Drugs and Technologies in Health (CADTH); 2016. URL: www.cadth.ca/sites/default/files/pdf/CP0015_PRESS_Update_Report_2016.pdf (accessed 1 January 2017).

142. Whiting P, Westwood M, Bojke L, Palmer S, Richardson G, Cooper J, *et al.* Clinical effectiveness and cost-effectiveness of tests for the diagnosis and investigation of urinary tract infection in children: a systematic review and economic model. *Health Technol Assess* 2006;**10**(36). https://doi.org/10.3310/hta10360

Appendix 1 Questionnaire

1. Please state your job title

_____

2. How long have you been searching databases such as MEDLINE (years)? _____

3. How often do you develop new search strategies as part of your work (For example searches to find treatments for conditions):

- None ☐
- Daily ☐
- Once a week ☐
- Once a month ☐
- Less than once a month ☐

4. What types of searches do you carry out (please tick all that apply):

- Rapid searches to answer brief queries ☐
- Scoping searches to estimate the size of the literature on a topic ☐
- Extensive searches to inform guidelines or systematic reviews ☐
- Other ☐

_____

_____

5. Which databases do you search regularly?

- MEDLINE ☐
- Embase ☐
- CINAHL ☐
- PSYCINFO ☐
- COCHRANE LIBRARY databases (CDSR, DARE, NHS EED, CENTRAL, HTA) ☐
- HEED ☐

_____

Please list any other databases that you use regularly in the box below

_____

_____

6. Methodological search filters (also known as Clinical Queries or Search Hedges) are used to find specific study designs such as randomized controlled trials. Have you ever used methodological search filters?

Yes ☐

No ☐

7. In what circumstances would you use methodological search filters?

Rapid searches to answer brief queries ☐

Scoping searches to estimate the size of the literature on a topic ☐

Extensive searches to inform guidelines or systematic reviews ☐

Other ☐

If Other, please describe below

_____

_____

8. Do you always use a filter when providing searches for similar types of projects? (for example, if you were searching for randomized controlled trials in MEDLINE would you always use a methodological search filter)?

Yes/No

☐

If No, please provide details about the circumstances when you would not use a filter)?

_____

_____

9. Please select the statement which describes your typical practice:

I use different search filters depending on whether my search has to be sensitive or precise ☐

I use the same search filter irrespective of the focus of the search ☐

10. If you had to find a methodological search filter for a specific study design where would you look?

_____

_____

11. What methodological search filters do you use at present?

Randomized controlled trials – please list the author or name of each of the filters you use?

_____

_____

Systematic reviews – please list the author or name of each of the filters you use?

_____

_____

Diagnostic studies – please list the author or name of each of the filters you use?

_____

_____

Studies of prognosis – please list the author or name of each of the filters you use?

_____

_____

Studies of etiology – please list the author or name of each of the filters you use?

_____

_____

Other trials – please list the author or name of each of the filters you use?

_____

_____

Guidelines – please list the author or name of each of the filters you use?

_____

_____

Economic evaluations – please list the author or name of each of the filters you use?

_____

_____

Other study methods – please list the author or name of each of the filters you use?

_____

_____

12. How do you decide which filter to use? Please select all which apply

Custom and practice – I've always used the same filters ☐
Guidance from a colleague ☐
I research the available filters and chose the best for my purposes ☐
I follow standing operating procedures/guidance on filters provided by my organization ☐
I use international/national guidance on best practice ☐
I use the filters available in the database interfaces I use e.g. Clinical Queries ☐

Please provide details on any other approaches you use to decide which filter to use.

_____

_____

13. Apart from adding a subject search, do you amend methodological search filters?

No ☐
Sometimes ☐
Always ☐

14. Please can you provide us with some more information about amending search filters?

Why, typically, do you amend search filters?

_____

How do you amend search filters?

_____

Do you test the effects of any amendments you make? **Yes/No**

If Yes, how do you test the amendments?

_____

Do you document the amendments when you write-up your searches? **Yes/No**

If Yes, how do you document the amendments?

_____

15. How do you keep up to date with methodological search filters? (Please tick all that apply)

Reading journal articles ☐

Current awareness services ☐

Please list typical current awareness services that you use

_____

Websites ☐

Please list typical websites that you use

_____

Professional development meetings and training events ☐

Email lists ☐

Please list typical email lists that you use

_____

RSS feeds ☐

Please list typical RSS feeds that you use

_____

Information provided by managers/work colleague ☐

Please use the box below to describe any other methods that you use to keep up to date with methodological search filters

16. If you have had to choose between methodological search filters what features or information has helped you to do so?

17. If you report your search process do you describe the filters you used? **Yes/No**

18. If you report your search process do you justify your choice of filters used? **Yes/No**

19. What do you think are the benefits of using methodological search filters?

20. What do you think are the limitations of using methodological search filters?

21. Imagine you have to choose between 2 or more methodological search filters:

What information would help you to choose which filter to use?

What would make choosing easier?

22. What methodological search filters would be useful to you?

23. Please use the box below to provide any further observations on methodological search filters as a tool for information retrieval.

Thank you for your help.

Appendix 2 Review C: search strategies and websites consulted that contained potentially relevant publications

Cochrane Methodology Register (The Cochrane Library, Issue 4 2011)

URL: www.thecochranelibrary.com/

Date searched: 18 October 2011.

Search strategy

#1 "diagnostic test accuracy":kw in Methods Studies

#2 "diagnostic test accuracy":kw and "search strategies".kw in Methods Studies

#3 (#1 AND NOT #2)

MEDLINE (1980 to October Week 3 2011), EMBASE (1980 to 2011 Week 43), MEDLINE In-Process & Other Non-Indexed Citations

Ovid Multifile Search: http://gateway.ovid.com/athens

Date searched: 18 October 2011.

Search strategy

1. *"diagnostic techniques and procedures"/ or *diagnostic imaging/ or *diagnostic tests, routine/ use mesz
2. *diagnostic accuracy/ or *diagnostic procedures/ or *Diagnostic test/ use emez
3. diagnostic.ti.
4. *roc curve/
5. *"sensitivity and specificity"/
6. or/1-5
7. *guidelines as topic/ use mesz
8. *practice guidelines/ use emez
9. *meta-analysis as topic/ use mesz
10. *meta-analysis/ use emez
11. *review literature as topic/ use mesz
12. *systematic review/ use emez
13. *Evidence-Based Medicine/mt, st use mesz
14. *evidence based medicine/ use emez
15. guideline?.ti.
16. (method$ adj1 standard$).ti.
17. methodological.ti.
18. (statistic$ adj1 method$).ti.
19. (working adj1 (party or committee or group)).ti.
20. or/7-19
21. 6 and 20

22. limit 21 to english language
23. remove duplicates from 22
24. limit 23 to yr="1980 –Current" (993)

Medion

Department of General Practice, University of Maastricht (www.mediondatabase.nl/)

Date searched: 18 October 2011.

Search: methodological studies on systematic reviews of diagnostic studies (all subheadings).

Websites consulted that contained potentially relevant publications

Date searched: 18 October 2011.

- AHRQ, US Department of Health and Human Services (www.ahrq.gov).
- Belgian Health Care Knowledge Centre (KCE) (www.kce.fgov.be/).
- CRD, University of York (www.york.ac.uk/inst/crd/).
- Diagnostic Test Accuracy Review Group, Cochrane (http://srdta.cochrane.org/welcome).
- Medical Services Advisory Committee, Australian Government Department of Health and Ageing (www.msac.gov.au/).
- NICE, Diagnostic Assessment Programme (www.nice.org.uk/aboutnice/whatwedo/aboutdiagnosticsassessment/diagnosticsassessmentprogramme.jsp).
- US FDA, US Department of Health and Human Services (www.fda.gov/).

Appendix 3 Review C: excluded studies

Arends LR, Hamza TH, van Houwelingen JC, Heijenbrok-Kal MH, Hunink MG, Stijnen T. Bivariate random effects meta-analysis of ROC curves. *Med Decis Making* 2008;**28**:621–38.

Begg CB. Methodologic standards for diagnostic test assessment studies. *J Gen Intern Med* 1988;**3**:518–20.

Bossuyt PM. Diagnostic accuracy reporting guidelines should prescribe reporting, not modeling. *J Clin Epidemiol* 2009;**62**:355–6, 362.

Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, *et al.* Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. *Ann Intern Med* 2003;**138**:40–4.

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. *Clin Chem* 2003;**49**:7–18.

Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, *et al.* Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. *Radiology* 2003;**226**:24–8.

Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, *et al.* Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. *Fam Pract* 2004;**21**:4–10.

Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, *et al.* Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. *Vet Clin Pathol* 2007;**36**:8–12.

Bruns DE, Huth EJ, Magid.E. Toward a checklist for reporting of studies of diagnostic accuracy of medical tests. *Clin Chem* 2000;**46**:893–5.

Chu H, Cole SR. Bivariate meta-analysis of sensitivity and specificity with sparse data: a generalized linear mixed model approach. *J Clin Epidemiol* 2006;**59**:1331–2, 13.

Chu H, Nie L, Cole SR, Poole C. Meta-analysis of diagnostic accuracy studies accounting for disease prevalence: alternative parameterizations and model selection. *Stat Med* 2009;**28**:2384–99.

Cleophas TJ, Droogendijk J, van Ouwerkerk BM. Validating diagnostic tests, correct and incorrect methods, new developments. *Curr Clin Pharmacol* 2008;**3**:70–6.

Elie C, Coste J. A methodological framework to distinguish spectrum effects from spectrum biases and to assess diagnostic and screening test accuracy for patient populations: application to the Papanicolaou cervical cancer smear test. *BMC Med Res Methodol* 2008;**8**:7.

Hamza TH, van Houwelingen HC, Heijenbrok-Kal MH, Stijnen T. Associating explanatory variables with summary receiver operating characteristic curves in diagnostic meta-analysis. *J Clin Epidemiol* 2009;**62**:1284–91.

Harbord RM, Deeks JJ, Egger M, Whiting P, Sterne JA. A unification of models for meta-analysis of diagnostic accuracy studies. *Biostatistics* 2007;**8**:239–51.

Harbord RM, Whiting P, Sterne JA, Egger M, Deeks JJ, Shang A *et al.* An empirical comparison of methods for meta-analysis of diagnostic accuracy showed hierarchical models are necessary. *J Clin Epidemiol* 2008;**61**:1095–103.

Hilden J. The area under the ROC curve and its competitors. *Med Decis Making* 1991;**11**:95–101.

Hollingworth W, Medina LS, Lenkinski RE, Shibata DK, Bernal B, Zurakowski D, *et al.* Interrater reliability in assessing quality of diagnostic accuracy studies using the QUADAS tool. A preliminary assessment. *Acad Radiol* 2006;**13**:803–10.

Irwig L. Modelling result-specific likelihood ratios. *J Clin Epidemiol* 1992;**45**:1335–8.

Irwig L, Tosteson AN, Gatsonis C, Lau J, Colditz G, Chalmers TC, *et al.* Guidelines for meta-analyses evaluating diagnostic tests. *Ann Intern Med* 1994;**120**:667–76.

Irwig L, Macaskill P, Glasziou P, Fahey M. Meta-analytic methods for diagnostic test accuracy. *J Clin Epidemiol* 1995;**48**:119–30.

Jaeschke R, Guyatt G, Sackett DL. Users' guides to the medical literature. III. How to use an article about a diagnostic test. A. Are the results of the study valid? Evidence-Based Medicine Working Group. *JAMA* 1994;**271**:389–91.

Jones CM, Athanasiou T. Diagnostic accuracy meta-analysis: review of an important tool in radiological research and decision making. *Br J Radiol* 2009;**82**:441–6.

Jones CM, Ashrafian H, Skapinakis P, Arora S, Darzi A, Dimopoulos K, *et al.* Diagnostic accuracy meta-analysis: a review of the basic principles of interpretation and application. *Int J Cardiol* 2010;**140**:138–44.

Khan KS. Systematic reviews of diagnostic tests: a guide to methods and application. *Best Pract Res Clin Obstet Gynaecol* 2005;**19**:37–46.

Knottnerus JA, van Weel C, Muris JWM. Evidence base of clinical diagnosis: evaluation of diagnostic procedures. *BMJ* 2002;**324**:477–80.

Lachs MS, Nachamkin I, Edelstein PH, Goldman J, Feinstein AR, Schwartz JS. Spectrum bias in the evaluation of diagnostic tests: lessons from the rapid dipstick test for urinary tract infection. *Ann Intern Med* 1992;**117**:135–40.

Leeflang M, Reitsma J, Scholten R, Rutjes A, Di Nisio M, Deeks J, *et al.* Impact of adjustment for quality on results of metaanalyses of diagnostic accuracy. *Clin Chem* 2007;**53**:164–72.

Lijmer JG, Bossuyt PM, Heisterkamp SH. Exploring sources of heterogeneity in systematic reviews of diagnostic tests. *Stat Med* 2002;**21**:1525–37.

Littenberg B, Moses LE. Estimating diagnostic accuracy from multiple conflicting reports: a new meta-analytic method. *Med Decis Making* 1993;**13**:313–21.

Lumbreras B, Porta M, Marquez S, Pollan M, Parker LA, Hernandez-Aguado I. QUADOMICS: an adaptation of the Quality Assessment of Diagnostic Accuracy Assessment (QUADAS) for the evaluation of the methodological quality of studies on the diagnostic accuracy of '-omics'-based technologies. *Clin Biochem* 2008;**41**:1316–25.

Lumbreras-Lacarra B, Ramos-Rincon JM, Hernandez-Aguado I. Methodology in diagnostic laboratory test research in *Clinical Chemistry* and *Clinical Chemistry and Laboratory Medicine*. *Clin Chem* 2004;**50**:530–6.

Macaskill P. Empirical Bayes estimates generated in a hierarchical summary ROC analysis agreed closely with those of a full Bayesian analysis. *J Clin Epidemiol* 2004;**57**:925–32.

Meads CA, Davenport CF. Quality assessment of diagnostic before–after studies: development of methodology in the context of a systematic review. *BMC Med Res Methodol* 2009;**9**:3.

Moher D, Tetzlaff J, Tricco AC, Sampson M, Altman DG. Epidemiology and reporting characteristics of systematic reviews. *PLOS Med* 2007;**4**:e78.

Mol BW, Lijmer JG, Evers JL, Bossuyt PM. Characteristics of good diagnostic studies. *Semin Reprod Med* 2003;**21**:17–25.

Mulherin SA, Miller WC. Spectrum bias or spectrum effect? Subgroup variation in diagnostic test evaluation. *Ann Intern Med* 2002;**137**:598–602.

Obuchowski NA. Sample size calculations in studies of test accuracy. *Stat Methods Med Res* 1998;**7**:371–92.

Oosterhuis WP, Niessen RW, Bossuyt PM. The science of systematic reviewing studies of diagnostic tests. *Clin Chem Lab Med* 2000;**38**:577–88.

Parker LA, Saez NG, Lumbreras B, Porta M, Hernandez-Aguado I. Methodological deficits in diagnostic research using '-omics' technologies: evaluation of the QUADOMICS tool and quality of recently published studies. *PLOS ONE* 2010;**5**:1–8.

Paul M, Riebler A, Bachmann LM, Rue H, Held L. Bayesian bivariate meta-analysis of diagnostic test studies using integrated nested Laplace approximations. *Stat Med* 2010;**29**:1325–39.

Petticrew MP, Sowden AJ, Lister SD, Wright K. False-negative results in screening programmes: systematic review of impact and implications. *Health Technol Assess* 2000;**4**(5).

Ransohoff D, Feinstein A. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *N Engl J Med* 1978;**299**:926–9.

Reid MC, Lachs MS, Feinstein AR. Use of methodological standards in diagnostic test research. Getting better but still not good. *JAMA* 1995;**274**:645–51.

Reynolds TA, Schriger DL. *Annals of Emergency Medicine* Journal Club. The conduct and reporting of meta-analyses of studies of diagnostic tests and a consideration of ROC curves: answers to the January 2010 Journal Club questions. *Ann Emerg Med* 2010;**55**:570–7.

Rigby AS, Summerton N. Statistical methods in epidemiology. VIII. On the use of likelihood ratios for diagnostic testing with an application to general practice. *Disabil Rehab* 2005;**27**:475–80.

Rutter CM, Gatsonis CA. Regression methods for meta-analysis of diagnostic test data. *Acad Radiol* 1995;**2**:S48–56.

Sackett DL, Haynes RB. The architecture of diagnostic research. *BMJ* 2002;**324**:539–41.

Schunemann HJ, Oxman AD, Brozek J, Glasziou P, Bossuyt P, Chang S, *et al.* GRADE: assessing the quality of evidence for diagnostic recommendations. *ACP J Club* 2008;**149**:2.

Schunemann HJ, Oxman AD, Brozek J, Glasziou P, Jaeschke R, Vist GE, *et al.* Grading quality of evidence and strength of recommendations for diagnostic tests and strategies. *BMJ* 2008;**336**:1106–10.

Schwenke C, Busse R. Analysis of differences in proportions from clustered data with multiple measurements in diagnostic studies. *Methods Inf Med* 2007;**46**:548–52.

Sheps SB, Schechter MT. The assessment of diagnostic tests: a survey of current medical research. *JAMA* 1984;**252**:2418–22.

Siadaty MS, Shu J. Proportional odds ratio model for comparison of diagnostic tests in meta-analysis. *BMC Med Res Methodol* 2004;**4**:27.

Smidt N, Overbeke J, De VH, Bossuyt P. Endorsement of the STARD statement by biomedical journals: survey of instructions for authors. *Clin Chem* 2007;**53**:1983–5.

Stengel D, Bauwens K, Sehouli J, Ekkernkamp A, Porzsolt F. A likelihood ratio approach to meta-analysis of diagnostic studies. *J Med Screen* 2003;**10**:47–51.

Suzuki S. Conditional relative odds ratio and comparison of accuracy of diagnostic tests based on 2 × 2 tables. *J Epidemiol* 2006;**16**:145–53.

Valenstein PN. Evaluating diagnostic tests with imperfect standards. *Am J Clin Pathol* 1990;**93**:252–8.

Walter SD, Irwig l, Glasziou PP. Meta-analysis of diagnostic tests with imperfect reference standards. *J Clin Epidemiol* 1999;**52**:943–51.

Appendix 4 Review D: search strategies

MEDLINE and MEDLINE In-Process & Other Non-Indexed Citations (OvidSP) (1950 to October Week 3 2010)

Date searched: 29 October 2010.

Search strategy

1. (methodolog$ adj3 filter$).ti,ab. (78)
2. (search adj3 filter$).ti,ab. (164)
3. (search adj strateg$).ti,ab. (9588)
4. (quality adj3 filter$).ti,ab. (278)
5. hedge$.ti,ab. (6400)
6. (clinical adj queries).ti,ab. (66)
7. ((economic or random$ or systematic or diagnostic) adj3 (filter? or search strateg$)).ti,ab. (618)
8. or/1-7 (16,583)
9. Choice Behavior/ (16,343)
10. (choice$ or choose or chose or choosing).ti,ab. (201,696)
11. select$.ti,ab. (944,947)
12. prefer$.ti,ab. (229,163)
13. (decid$ or decision$).ti,ab. (190,433)
14. judgment$.ti,ab. (21,984)
15. or/9-14 (1,468,055)
16. 8 and 15 (8714)
17. Librarians/ (600)
18. librarian$.ti,ab. (1773)
19. (information adj2 (specialist$ or officer$ or scientist$)).ti,ab. (474)
20. (searcher$ or researcher$).ti,ab. (56,147)
21. or/17-20 (58,457)
22. 16 and 21 (638)

EMBASE (OvidSP) (1980 to Week 42 2010)

Date searched: 29 October 2010.

Search strategy

1. information retrieval/ and methodology/ (6053)
2. (methodolog$ adj3 filter$).ti,ab. (93)
3. (search adj3 filter$).ti,ab. (189)
4. (search adj strateg$).ti,ab. (11,533)
5. (quality adj3 filter$).ti,ab. (370)
6. hedge$.ti,ab. (6951)
7. (clinical adj queries).ti,ab. (75)
8. ((economic or random$ or systematic or diagnostic) adj3 (filter? or search strateg$)).ti,ab. (709)
9. or/1-8 (25,162)
10. decision making/ (101,825)
11. (choice$ or choose or chose or choosing).ti,ab. (243,717)
12. select$.ti,ab. (1,095,085)

13. prefer$.ti,ab. (256,127)
14. (decid$ or decision$).ti,ab. (226,199)
15. judgment$.ti,ab. (24,055)
16. or/10-15 (1,746,266)
17. 9 and 16 (11,520)
18. librarian/ (736)
19. librarian$.ti,ab. (1650)
20. (information adj2 (specialist$ or officer$ or scientist$)).ti,ab. (553)
21. (searcher$ or researcher$).ti,ab. (65,524)
22. or/18-21 (67,855)
23. 17 and 22 (824)

PsycINFO (OvidSP) (1806 to October Week 4 2010)

Date searched: 29 October 2010.

Search strategy

1. (methodolog$ adj3 filter$).ti,ab. (9)
2. (search adj3 filter$).ti,ab. (29)
3. (search adj strateg$).ti,ab. (1062)
4. (quality adj3 filter$).ti,ab. (19)
5. hedge$.ti,ab. (620)
6. (clinical adj queries).ti,ab. (5)
7. ((economic or random$ or systematic or diagnostic) adj3 (filter? or search strateg$)).ti,ab. (65)
8. or/1-7 (1743)
9. choice behavior/ (11,420)
10. (choice$ or choose or chose or choosing).ti,ab. (102,121)
11. select$.ti,ab. (190,035)
12. prefer$.ti,ab. (81,835)
13. (decid$ or decision$).ti,ab. (112,977)
14. judgment$.ti,ab. (49,128)
15. or/9-14 (460,385)
16. 8 and 15 (521)
17. exp information specialists/ (174)
18. librarian$.ti,ab. (515)
19. (information adj2 (specialist$ or officer$ or scientist$)).ti,ab. (176)
20. (searcher$ or researcher$).ti,ab. (72,543)
21. or/17-20 (73,226)
22. 16 and 21 (30)

Library, Information Science and Technology Abstracts (LISTA) (EBSCO*host*) (1986 to October 2010)

Date searched: 29 October 2010.

Search strategy

S22 S15 and S21 (164)

S21 S16 or S17 or S18 or S19 or S20 (118,186)

S20 TX searcher* or researcher* (14,140)

S19 TX information N2 specialist* or information N2 officer* or information N2 scientist* (5812)

S18 TX librarian* (102,368)

S17 DE "INFORMATION professionals" (3235)

S16 DE "LIBRARIANS" (18,635)

S15 S11 and S14 (468)

S14 S12 or S13 (122,568)

S13 TX select* or prefer* or decid* or decision* or judgment* (56,972)

S12 TX choice* or choose or chose or choosing (68,911)

S11 S1 or S2 or S3 or S4 or S5 or S6 or S7 or S8 or S9 or S10 (2074)

S10 TX diagnostic N3 filter? or random* N3 search* (61)

S9 TX systematic N3 filter? or random* N3 search* (62)

S8 TX random* N3 filter? or random* N3 search (50)

S7 TX economic N3 filter? or economic N3 search* (55)

S6 TX "clinical queries" (20)

S5 TX hedge* (423)

S4 TX quality N3 filter* (54)

S3 TX search N1 strateg* (1400)

S2 TX search N3 filter* (106)

S1 TX methodolog* N3 filter* (13)

Cochrane Methodology Register (The Cochrane Library) (Issue 4 2010)

URL: www.thecochranelibrary.com/

Date searched: 29 October 2010.

Search strategy

#1 (methodolog* NEAR/3 filter*):ti,ab,kw (30)

#2 (search NEAR/3 filter*):ti,ab,kw (85)

#3 (search NEXT strateg*):ti,ab,kw (5136)

#4 (quality NEAR/3 filter*):ti,ab,kw (20)

#5 (hedge*):ti,ab,kw (32)

#6 (clinical NEXT queries):ti,ab,kw (24)

#7 (economic or random* or systematic or diagnostic) NEAR/3 (filter? or search strateg*):ti,ab,kw (236)

#8 (#1 OR #2 OR #3 OR #4 OR #5 OR #6 OR #7) (5214)

#9 MeSH descriptor Choice Behavior explode all trees (696)

#10 (choice* or choose or chose or choosing):ti,ab,kw (1727)

#11 (select*):ti,ab,kw (5287)

#12 (prefer*):ti,ab,kw (11,236)

#13 (decid* or decision*):ti,ab,kw (10,668)

#14 (judgment*):ti,ab,kw (1335)

#15 (#9 OR #10 OR #11 OR #12 OR #13 OR #14) (70,646)

#16 (#8 AND #15) (4704)

#17 MeSH descriptor Librarians explode all trees (5)

#18 (librarian*):ti,ab,kw (144)

#19 (information NEAR/2 (specialist* or officer* or scientist*)):ti,ab,kw (34)

#20 (searcher* or researcher*):ti,ab,kw (2534)

#21 (#17 OR #18 OR #19 OR #20) (2677)

#22 (#16 AND #21) (458)

Science Citation Index (1899–2010), Social Science Citation Index (1956–2010), Conference Proceedings Citation Index – Science (1990–2010) and Conference Proceedings Citation Index – Social Science and Humanities (1990–2010) (ISI Web of Science)

Search date: 29 October 2010.

Search strategy

#18 420 #13 and #17

Databases=SCI-EXPANDED Timespan=All Years

#17 71,421 #14 or #15 or #16

Databases=SCI-EXPANDED Timespan=All Years

#16 66,699 TS=(searcher* or researcher*)

Databases=SCI-EXPANDED Timespan=All Years

#15 2970 TS=(information) SAME TS=(specialist* or officer* or scientist*)

Databases=SCI-EXPANDED Timespan=All Years

#14 2057 TS=librarian*

Databases=SCI-EXPANDED Timespan=All Years

#13 8269 #8 and #12

Databases=SCI-EXPANDED Timespan=All Years

#12 > 100,000 #9 or #10 or #11

Databases=SCI-EXPANDED Timespan=All Years

#11 20,409 TS=judgment*

Databases=SCI-EXPANDED Timespan=All Years

#10 > 100,000 TS=(select* or prefer* or decid* or decision*)

Databases=SCI-EXPANDED Timespan=All Years

#9 > 100,000 TS=(choice* or choose or chose or choosing)

Databases=SCI-EXPANDED Timespan=All Years

#8 30,806 #1 or #2 or #3 or #4 or #5 or #6 or #7

Databases=SCI-EXPANDED Timespan=All Years

#7 5102 TS=(economic or random* or systematic or diagnostic) SAME TS=(filter* or search strateg*)

Databases=SCI-EXPANDED Timespan=All Years

#6 46 TS=("clinical queries")

Databases=SCI-EXPANDED Timespan=All Years

#5 14,192 TS=hedge*

Databases=SCI-EXPANDED Timespan=All Years

#4 3183 TS=(quality SAME filter*)

Databases=SCI-EXPANDED Timespan=All Years

#3 7524 TS=("search strateg*")

Databases=SCI-EXPANDED Timespan=All Years

#2 1102 TS=(search SAME filter*)

Databases=SCI-EXPANDED Timespan=All Years

#1 814 TS=(methodology* SAME filter*)

Databases=SCI-EXPANDED Timespan=All Years

Health Techology Assessment international Vortal

URL: www.htai.org/index.php?id=577

Date searched: 29 October 2010.

Search strategy

"methodological filter" choice librarian
"methodological filter" choice specialist
"methodological filter" choice searcher
"methodological filter" choice researcher
"methodological filter" choice officer
"methodological filter" decide librarian
"methodological filter" decide specialist
"methodological filter" decide searcher
"methodological filter" decide researcher
"methodological filter" decide officer

"search filter" choice librarian
"search filter" choice specialist
"search filter" choice searcher
"search filter" choice researcher
"search filter" choice officer
"search filter" decide librarian
"search filter" decide specialist
"search filter" decide searcher
"search filter" decide researcher
"search filter" decide officer

"search strategy" choice librarian
"search strategy" choice "information specialist"
"search strategy" choice searcher
"search strategy" choice "information officer"
"search strategy" decide librarian
"search strategy" decide "information specialist"
"search strategy" decide searcher
"search strategy" decide "information officer"

View all resources

Searching the HTA Literature
MEDLINE/PubMed
Clinical Trial Registries

Evaluated Sources
Grey Literature
Information on Literature Searching
Searching on the Web
Clinical Practice Guidelines

Reference tools
Keeping Up: stuff for Librarians and Information Specialists

European network for Health Technology Assessment

URL: www.eunethta.net/

Date searched: 1 November 2010.

General search
+search +filter
+methodological +filter
+search +strategy
+search +strategies

Tools
EUnetHTA Planned and Ongoing Projects (POP) Database
EUnetHTA Database on additional evidence
EUnetHTA News Aggregator
HTA Core Model

All resources require username and password to access – EUnetHTA membership (partners and associates)

Health technology assessment organisation websites

Date searched: 1–3 November 2010.

- Agencia de Evaluación de Tecnologías Sanitarias (AETS) (www.isciii.es/htdocs/en/investigacion/Agencia_quees.jsp).
- AHRQ (www.ahrq.gov/).
- Basque Office for Health Technology Assessment (OSTEBA) (www.osanet.euskadi.net/osteba/es).
- CADTH (www.cadth.ca/index.php/en/home).
- CRD (www.york.ac.uk/inst/crd/).
- Comité d'Evaluation et de Diffusion des Innovations Technologiques (CEDIT) (http://cedit.aphp.fr/).
- German Agency for HTA at the German Institute for Medical Documentation and Information (DAHTA@DIMDI) (www.dimdi.de).
- Institute for Quality and Efficiency in Health Care (IQWiG) (www.iqwig.de/).
- International Network of Agencies for Health Technology Assessment (INAHTA) (www.inahta.org/).
- Swedish Council on Health Technology Assessment (SBU) (www.sbu.se/en/).

Health Libraries Group

URL: www.cilip.org.uk/get-involved/special-interest-groups/health/Pages/default.aspx

Date searched: 1 November 2010.

Search this group

"search filter"
"methodological filter"
"search strategy"
"search strategies"
hedges

Search the IFM Healthcare website (www.ifmh.org.uk/)

"search filter"
"methodological filter"
"search strategy"
"search strategies"
hedges

European Association for Health Information and Libraries

URL: www.eahil.net/

Date searched: 1 November 2010.

Search

"search filter"
"methodological filter"
"search strategy"
"search strategies"
Hedges

US Medical Library Association

URL: www.mlanet.org/

Date searched: 1 November 2010.

Search

"search filter"
"methodological filter"
"search strategy"
"search strategies"
Hedges

Appendix 5 Review E: search strategies

EMBASE (1980 to 2011 Week 9), Ovid MEDLINE(R) (1948 to February Week 4 2011), Ovid MEDLINE(R) In-Process & Other Non-Indexed Citations (8 March 2011)

Ovid Multifile Search: https://shibboleth.ovid.com/

Date searched: 8 March 2011.

Search strategy

1. choice behavior/ use mesz
2. decision making/
3. professional practice/
4. physician's practice patterns/ use mesz
5. clinical practice/ use emez
6. ((clinician$ or physician$ or doctor$ or practitioner$) adj3 (choice$ or chos$ or choos$)).ti.
7. ((clinician$ or physician$ or doctor$ or practitioner$) adj3 (select$ or decid$ or decision$)).ti.
8. ((clinician$ or physician$ or doctor$ or practitioner$) adj3 prefer$).ti.
9. 1 or 2
10. 3 or 4 or 5
11. 9 and 10
12. 6 or 7 or 8 or 11
13. exp "diagnostic techniques and procedures"/ use mesz
14. exp diagnosis/ use emez
15. (diagnosis or diagnostic$).ti,hw.
16. (test or tests).ti,hw.
17. 13 or 14 or 15 or 16
18. 12 and 17
19. remove duplicates from 18
20. limit 19 to english
21. (abstract or comment or conference or letter).pt
22. 20 not 21

PsycINFO (1987 to 8 June 2011)

EBSCO*host* (http://web.ebscohost.com/)

Date searched: 8 June 2011.

Search strategy

S1 DE "Choice Behavior" OR DE "Decision Making"

S2 TX clinician* n3 prefer* or TX physician* n3 prefer* or TX doctor* n3 prefer* or TX practitioner* n3 prefer*

S3 TX clinician* n3 decision* or TX physician* n3 decision* or TX doctor* n3 decision* or TX practitioner* n3 decision*

S4 TX clinician* n3 decid* or TX physician* n3 decid* or TX doctor* n3 decid* or TX practitioner* n3 decid*

S5 TX clinician* n3 select* or TX physician* n3 select* or TX doctor* n3 select* or TX practitioner* n3 select*

S6 TX clinician* n3 choos* or TX physician* n3 choos* or TX doctor* n3 choos* or TX practitioner* n3 choos*

S7 TX clinician* n3 chos* or TX physician* n3 chos* or TX doctor* n3 chos* or TX practitioner* n3 chos*

S8 TX clinician* n3 choice* or TX physician* n3 choice* or TX doctor* n3 choice* or TX practitioner* n3 choice*

S9 S1 or S2 or S3 or S4 or S5 or S6 or S7 or S8

S10 DE "Screening" OR DE "Health Screening"

S11 DE "Diagnosis" OR DE "Differential Diagnosis" OR DE "Medical Diagnosis"

S12 TX test* n3 order* or TX diagnos* n3 test* or TX screen* n3 test*

S13 S10 or S11 or S12

S14 S9 and S13 Limiters - English

Cumulative Index to Nursing and Allied Health Literature (1983 to 9 June 2011)

EBSCO*host* (http://web.ebscohost.com/)

Date searched: 9 June 2011.

Search strategy

S1 (MM "Decision Making")

S2 (MM "Practice Patterns") OR (MM "Professional Practice")

S3 TX clinician* n3 prefer* or TX physician* n3 prefer* or TX doctor* n3 prefer* or TX practitioner* n3 prefer*

S4 TX clinician* n3 decision* or TX physician* n3 decision* or TX doctor* n3 decision* or TX practitioner* n3 decision*

S5 TX clinician* n3 decid* or TX physician* n3 decid* or TX doctor* n3 decid* or TX practitioner* n3 decid*

S6 TX clinician* n3 select* or TX physician* n3 select* or TX doctor* n3 select* or TX practitioner* n3 select*

S7 TX clinician* n3 choos* or TX physician* n3 choos* or TX doctor* n3 choos* or TX practitioner* n3 choos*

S8 TX clinician* n3 chos* or TX physician* n3 chos* or TX doctor* n3 chos* or TX practitioner* n3 chos*

S9 TX clinician* n3 choice* or TX physician* n3 choice* or TX doctor* n3 choice* or TX practitioner* n3 choice* Search modes - Boolean/Phrase

S10 S1 or S2 or S3 or S4 or S5 or S6 or S7 or S8 or S9

S11 TX test* n3 order* or TX diagnos* n3 test* or TX screen* n3 test* -

S12 MW diagnosis

S13 (MH "Diagnosis")

S14 (MH "Health Screening+")

S15 S11 or S12 or S13 or S14

S16 S10 and S15 Limiters - English Language

Applied Social Sciences Index and Abstracts (1987 to 13 June 2011)

CSA Illumina (www.csa.com/)

Date searched: 13 June 2011.

Search strategy

Search Query #29 (((DE = choice) or(DE = clinical decision making) or(DE = clinical practice) or (TI = ((clinician* or physician* or doctor* or practitioner*) within 3 (choice* or chos* or choos*))) or (AB = ((clinician* or physician* or doctor* or practitioner*) within 3 (choice* or chos* or choos*))) or(TI = ((clinician* or physician* or doctor* or practitioner*) within 3 (select* or decid* or decision*))) or (AB = ((clinician* or physician* or doctor* or practitioner*) within 3 (select* or decid* or decision*))) or(TI = ((clinician* or physician* or doctor* or practitioner*) within 3 (prefer*))) or(AB = ((clinician* or physician* or doctor* or practitioner*) within 3 (prefer*)))) and((DE = diagnostic testing) or(TI = (diagnosis or diagnostic* or test or tests)))) or(TI = (test* within 3 order* within 3 (choice* or chos* or choos*))) or (AB = (test* within 3 order* within 3 (choice* or chos* or choos*))) or(AB = (test* within 3 order*) AND (choice* or chos* or choice*)) or(TI = (test* within 3 order*) AND (choice* or chos* or choice*)) or(TI = (diagnos* within 3 test*) AND (choice* or chos* or choice*)) or(AB = (diagnos* within 3 test*) AND (choice* or chos* or choice*)) or(AB = (diagnos* within 3 test*) AND (select* or decid* or decision*)) or (TI = (diagnos* within 3 test*) AND (select* or decid* or decision*)) or(TI = (test* within 3 order*) AND (select* or decid* or decision*)) or(AB = (test* within 3 order*) AND (select* or decid* or decision*)) or (AB = (test* within 3 order*) AND (prefer*)) or(TI = (test* within 3 order*) AND (prefer*)) or(TI = (diagnos* within 3 test*) AND (prefer*)) or(AB = (diagnos* within 3 test*) AND (prefer*))

National screening programmes (accessed July 2011)

- Australian Population Health Development Screening Subcommittee (www.health.gov.au/internet/screening/publishing.nsf/Content/home).
- UK National Screening Committee (www.screening.nhs.uk/).
- US Preventive Services Task Force (www.ahrq.gov/clinic/uspstfix.htm).
- World Health Organization (www.who.int/).

Appendix 6 Review E: excluded studies

Record	Exclusion reason
***Diagnostic reasoning* (n = *10*)**	
Bornstein BH, Emler AC. Rationality in medical decision making: a review of the literature on doctors' decision-making biases. *J Eval Clin Pract* 2001;**7**:97–107	Reviews biases that result in suboptimal diagnostic decisions
Cahan A, Gilon D, Manor O, Paltiel O. Probabilistic reasoning and clinical decision-making: do doctors overestimate diagnostic probabilities? *Q J Med* 2003;**96**:763–9	Subadditivity in physicians' estimates of pretest probability
Croskerry P. Diagnostic failure: a cognitive and affective approach. *Adv Patient Safety* 2004;**2**:241–54	Factors leading to errors in diagnostic reasoning
Hays DG, McLeod AL, Prosek E. Diagnostic variance among counselors and counselor trainees. *Measure Eval Counsel Develop* 2009;**42**:3–14	Variance in diagnostic reasoning
Heller R, Sandars JE, Patterson L, McElduff P. GPs' and physicians' interpretation of risks, benefits and diagnostic test results. *Fam Pract* 2004;**21**:155–9	Physicians' understanding of pretest probability and baseline risk and application to diagnostic test results
Klein JG. Five pitfalls in decisions about diagnosis and prescribing. *BMJ* 2005;**330**:781–3	Errors in diagnostic reasoning
Lutfey KE, Link CL, Marceau LD, Grant RW, Adams A, Arber S, *et al.* Diagnostic certainty as a source of medical practice variation in coronary heart disease: results from a cross-national experiment of clinical decision making. *Med Decis Making* 2009;**29**:606–18	Diagnostic certainty influence on patient management, including test ordering
Sassi F, McKee M. Do clinicians always maximize patient outcomes? A conjoint analysis of preferences for carotid artery testing. *J Health Serv Res Policy* 2008;**13**:61–6	Conjoint analysis to elicit how physicians value different diagnostic test characteristics
Shemberg KM, Doherty ME. Is diagnostic judgment influenced by a bias to see pathology? *J Clin Psychol* 1999;**55**:513–18	Biases in diagnostic reasoning
Steurer J, Fischer JE, Bachmann LM, Koller M, ter Riet G. Communicating accuracy of tests to general practitioners: a controlled study. *BMJ* 2002;**324**:824–6	Physicians' understanding of diagnostic accuracy statistics and how presentation of test results influences estimates of disease probability
***Test use* (n = *10*)**	
Charles RF, Powe NR, Jaar BG, Troll MU, Parekh RS, Boulware LE. Clinical testing patterns and cost implications of variation in the evaluation of CKD among US physicians. *Am J Kidney Dis* 2009;**54**:227–37	Survey of chronic kidney disease clinical practice guideline adherence including test use
Cleary-Goldman J, Morgan MA, Malone FD, Robinson JN, D'Alton ME, Schulkin J. Screening for Down syndrome: practice patterns and knowledge of obstetricians and gynecologists. *Obstet Gynecol* 2006;**107**:11–17	Practice patterns on screening for Down syndrome
Gringas P. Choice of medical investigations for developmental delay: a questionnaire survey. *Child Care Health Develop* 1998;**24**:267–76	Survey on diagnostic test use for developmental delay
Kitahara S, Iwatsubo E, Yasuda K, Ushiyama T, Nakai H, Suzuki T, *et al.* Practice patterns of Japanese physicians in urologic surveillance and management of spinal cord injury patients. *Spinal Cord* 2006;**44**:362–8	Survey on test use for urological surveillance in spinal cord injury patients
Mangat J, Conron M, Gabbay E, Proudman SM; Pulmonary Interstitial Vascular Organisational Taskforce (PIVOT). Scleroderma lung disease, variation in screening, diagnosis and treatment practices between rheumatologists and respiratory physicians. *Intern Med J* 2010;**40**:494–502	Compares management of scleroderma lung disease between specialities
McGregor SE, Hilsden RJ, Murray A, Bryant HE. Colorectal cancer screening: practices and opinions of primary care physicians. *Prev Med* 2004;**39**:279–85	Survey on adherence to national guidelines for colorectal cancer screening

Record	Exclusion reason
Oxentenko AS, Vierkant RA, Pardi DS, Farley DR, Dozois EJ, Hartman TE, *et al.* Colorectal cancer screening perceptions and practices: results from a national survey of gastroenterology, surgery and radiology trainees. *J Cancer Educ* 2007;**22**:219–26	Survey of perceptions of different tests for colorectal cancer
Plaut D. A committee approach to test utilization. *AMT Events* 2010;**27**:164–5	Guidance on optimising laboratory test use
Spiegel BM, Ho W, Esrailian E, Targan S, Higgins PDR, Siegel CA, *et al.* Controversies in ulcerative colitis: a survey comparing decision making of experts versus community gastroenterologists. *Clin Gastroenterol Hepatol* 2009;**7**:168–74	Survey on management of Crohn's disease, including test use
You JJ, Levinson W, Laupacis A. Attitudes of family physicians, specialists and radiologists about the use of computed tomography and magnetic resonance imaging in Ontario. *Healthcare Policy* 2009;**5**:54–65	Survey on computerised tomography and magnetic resonance imaging use
***Diagnostic process/strategy* (n = *6*)**	
Eken C, Ercetin Y, Ozgurel T, Kilicaslan Eray O. Analysis of factors affecting emergency physicians' decisions in the management of chest pain patients. *Eur J Emerg Med* 2006;**13**:214–17	Factors that affect physicians' decisions in the diagnosis of patients with chest pain
Fischer T, Fischer S, Himmel W, Kochen MM, Hummer-Pradier E. Family practitioners' diagnostic decision-making processes regarding patients with respiratory tract infections: an observational study. *Med Decis Making* 2008;**28**:810–18	Physicians' diagnostic strategies for patients with respiratory tract infection symptoms
Roy JS, Michlovitz S. Using evidence-based practice to select diagnostic tests. *Hand Clin* 2009;**25**:49–57	Benefits of using evidence-based practice to improve diagnostic test selection
Salkeld EJ. Integrative medicine and clinical practice: diagnosis and treatment strategies. *Complement Health Pract Rev* 2008;**13**:21–33	Use of complementary and traditional medicine in diagnostic strategies
von dem Knesebeck, Bönte M, Siegrist J, Marceau L, Link C, Arber S, *et al.* Country differences in the diagnosis and management of coronary heart disease – a comparison between US, UK and Germany. *BMC Health Serv Res* 2008;**8**:198	The impact of structural issues on diagnostic processes
Whiting P, Toerien M, de Salis I, Sterne JA, Dieppe P, Egger M, *et al.* A review identifies and classifies reasons for ordering diagnostic tests. *J Clin Epidemiol* 2007;**60**:981–9	Review factors that influence test ordering decisions
***One test choice* (n = *5*)**	
Baker SR, Susman PH, Sheen L, Pan L. Comparison of test-ordering choices of college physicians and emergency physicians for young adults with abdominal pain: influences and preferences for CT use. *Emerg Radiol* 2010;**17**:455–9	Computerised tomography scanning for two clinical scenarios
Espeland A, Baerheim A. Factors affecting general practitioners' decisions about plain radiography for back pain: implications for classification of guideline barriers – a qualitative study. *BMC Health Serv Res* 2003;**3**:8	Factors that influence the decision to order radiography for back pain
Haggerty JT, Tudiver F, Brown JB, Herbert C, Ciampi A, Guibert R, *et al.* Patients' anxiety and expectations: how they influence family physicians' decisions to order cancer screening tests. *Can Fam Physician* 2005;**51**:1658–9	Factors that influence the decision to order screening tests
Lewis JD, Asch DA, Ginsberg GG,Hoops TC, Kochman ML, Bilker WB, Strom BL. Primary care physicians' decisions to perform flexible sigmoidoscopy. *J Gen Intern Med* 1999;**14**:297–302	Factors that influence physicians' decision to order flexible sigmoidoscopy
Szeinbach SL, Harpe SE, Williams PB, Elhefni H. Testing for allergic disease: parameters considered and test value. *BMC Health Serv Res* 2008;**9**:47	Factors that influence the decision to order a blood test for allergic rhinitis
***Patient choice/compliance* (n = *4*)**	
Heckerling PS, Verp MS, Albert N. The role of physician preferences in the choice of amniocentesis or chorionic villus sampling for prenatal genetic testing. *Genet Test* 1998;**2**:61–6	Effect of physician characteristics, including preferences, on patient choice

Record	Exclusion reason
Heckerling PS, Verp MS, Albert N. Patient or physician preferences for decision analysis: the prenatal genetic testing decision. *Med Decis Making* 1999;**19**:66–77	Decision analysis of patient and physician preferences to predict patient choice
Marshall DA, Johnson FR, Kulin NA, Ozdemir S, Walsh JM, Marshall JK, *et al.* How do physician assessments of patient preferences for colorectal cancer screening tests differ from actual preferences? A comparison in Canada and the United States using a stated-choice survey. *Health Econ* 2009;**18**:1420–39	Stated preferences discrete choice survey of patients and physicians on patient preferences for colorectal cancer screening tests
Murphy DJ, Gross R, Buchanan J. Computerized reminders for five preventive screening tests: generation of patient-specific letters incorporating physician preferences. *Proc AMIA Symp* 2000;600–4	Effect of computer reminders on attendance for screening
***Interventions to influence test ordering* (n = *2*)**	
Hampers LC, Cha S, Gutglass DJ, Krug SE, Binns HJ. The effect of price information on test-ordering behavior and patient outcomes in a pediatric emergency department. *Pediatrics* 1999;**103**:877–82	Effect of price information on test ordering
Kashner T, Rush AJ, Surís A, Biggs MM, Gajewski VL, Hooker DJ, *et al.* Impact of structured clinical interviews on physicians' practices in community mental health settings. *Psychiatr Serv* 2003;**54**:712–18	Effect on disease management, including test ordering, of providing physicians with the results of clinical interviews
***Test choice but reasons not obtained* (n = *2*)**	
Carey TS, Garrett J. Patterns of ordering diagnostic tests for patients with acute low back pain. *Ann Intern Med* 1996;**125**:807–14	Survey of factors associated with test choice
Pereira B, Tamer M, Khalifa K, Mokbel K, *et al.* General practitioners' greater choice for sentinel node biopsy that patients in the UK. *Curr Med Res Opinion* 2004;**20**:417–18	Physicians' preferences for biopsy test
***Economic model* (n = *1*)**	
Vijan S, Hwang EW, Hofer TP, Hayward RA. Which colon cancer screening test? A comparison of cost, effectiveness and compliance. *Am J Med* 2001;**111**:593–601	Cost-effectiveness of different screening strategies for colon cancer

EME
HS&DR
HTA
PGfAR
PHR

Part of the NIHR Journals Library
www.journalslibrary.nihr.ac.uk

Published by the NIHR Journals Library