Evaluation of PubMed filters used for evidence-based searching: validation using relative recall

Arjen Hoogendam, MD; Pieter F. de Vries Robbé, MD, PhD; Anton F. H. Stalenhoef, MD, PhD, FRCP; A. John P. M. Overbeke, MD, PhD

See end of article for authors' affiliations.

DOI: 10.3163/1536-5050.97.3.007

Objectives: The research sought to determine the value of PubMed filters and combinations of filters in literature selected for systematic reviews on therapy-related clinical questions.

Methods: References to 35,281 included and 48,514 excluded articles were extracted from 2,629 reviews published prior to January 2008 in the Cochrane Database of Systematic Reviews and sent to PubMed with and without filters. Sensitivity, specificity, and precision were calculated from the percentages of unfiltered and filtered references retrieved for each review and averaged over all reviews.

Results: Sensitivity of the Sensitive Clinical Queries filter was reasonable (92.7%, 92.1–93.3); specificity (16.1%, 15.1–17.1) and precision were low (49.5%, 48.5–50.5). The Specific Clinical Queries and the Single Term Medline Specific filters performed comparably (sensitivity, 78.2%, 77.2–79.2 vs. 78.0%; 77.0–79.0; specificity, 52.0%, 50.8–53.2 vs. 52.3%, 51.1–53.5; precision, 60.4%, 59.4–61.4 vs. 60.6%, 59.6–61.6). Combining the Abridged Index Medicus (AIM) and Single Term Medline Specific (65.2%, 63.8–66.6), Two Terms Medline Optimized (64.2%, 62.8–65.6), or Specific Clinical Queries filters (65.0%, 63.6–66.4) yielded the highest precision.

Conclusions: Sensitive and Specific Clinical Queries filters used to answer questions about therapy will result in a list of clinical trials but cannot be expected to identify only methodologically sound trials. The Specific Clinical Queries filters are not suitable for questions regarding therapy that cannot be answered with randomized controlled trials. Combining AIM with specific PubMed filters yields the highest precision in the Cochrane dataset.

INTRODUCTION

PubMed [1] is one of the major sources of medical information. Although information sources containing integrated data, like online textbooks and guidelines, are more practical for handling daily clinical questions, PubMed or comparable databases are indispensable for answering detailed questions, finding information on rare diseases, or uncovering the latest developments [2]. Physicians trying to answer patient-related questions using PubMed during daily medical care are confronted with the difficult task of retrieving only relevant information. Retrieving a limited set of articles that is likely to answer the question requires skill. After potentially relevant articles have been retrieved, a critical appraisal must follow to determine the methodological quality of each study. Tools that help in retrieving a small set of methodologically strong trials can help the physician find relevant answers to the question.

Filters help reduce the number of articles retrieved by selecting articles based on specific characteristics. PubMed provides single term filters, available from the limits section, that select articles based on specific Medical Subject Headings (MeSH), publication types, or dates to narrow the search. Some of these limits are recommended by evidence-based search guidelines as being particularly suited to answering patient-related questions [3, 4]. Identification of reports based on methodologically sound trials, which comply with internationally accepted rules for conducting scientific trials and reporting results (Consolidated Standards of Reporting Trials [CONSORT] Statement [5]), can help reduce the number of irrelevant results. Haynes et al. developed special search filters aimed at retrieving methodologically sound trials about therapy, diagnosis, prognosis, and etiology [6–11]. To develop these filters, Haynes et al. used a set of 161 clinical journals [7] from which all articles were evaluated for relevance to the subject (therapy, diagnosis, etc.) and appraised for methodological quality according to the process delineated in Table 1.

Haynes et al. designed both sensitive and specific filters regarding therapy using this method. The sensitivity of sensitive filters, aimed at retrieving all the relevant information, and the specificity of specific filters, aimed at correctly identifying irrelevant information, are very high, suggesting that they are fully able to select methodologically strong trials regarding therapy. As in the design of diagnostic tests, sensitivity and specificity are markers for the quality of a filter (Table 2).

More relevant to physicians solving medical questions, however, is the precision of a filter that determines the percentage of methodologically sound, relevant articles in the complete set of articles that a query retrieves. The amount of time needed to find a relevant article in a large set of irrelevant articles is likely to exceed the time available during daily medical care [2, 12–15]. Reducing the number of articles to read using precise filters is therefore crucial for on-the-spot searching. The precision of the Clinical Queries as reported by Haynes et al. is low, ranging from 10%–60% [7]. They suggested that it might even be lower when the filters are used in PubMed searches.

Highlights

- Sensitivity and specificity of PubMed filters are low when used in a set of articles reporting clinical trials relevant to questions about therapy.
- The precision of Clinical Queries filter was higher in a set of articles reporting clinical trials relevant to questions about therapy than previously reported in a set not restricted to questions about therapy, but precision was still too low to be of practical use in daily medicine.
- The highest single filter precision was achieved by the Single Term Medline Specific filter, which uses ''randomized Controlled Trial[ptyp]'' as its filter criterion.
- Combining the Abridged Index Medicus filter with the Specific Clinical Queries, Two Terms Medline Optimized, or the Single Term Medline Specific filters resulted in the highest precision.
- Filtering for Humans and English had low precision in a set of articles reporting clinical trials relevant to questions about therapy.

Implications

- Sensitive and Specific Clinical Queries filters may be used to answer questions about therapy and will result in the selection of (randomized) controlled trials, but the filters cannot be expected to identify only methodologically sound trials.
- Adequately formed queries for therapeutic questions have an advantage compared with Clinical Queries filters in that the physician can search for more study types than only trials and will retrieve articles about subjects that cannot be studied with randomized controlled trials.
- PubMed filters should be validated by external validation, preferably by combining them with user queries to establish their real value for on-the-spot searching.

As filters are used in combination with queries, precision will depend strongly on the quality of the query. It is likely that filters are much more precise when used with a set of articles retrieved in respect to a specific clinical question, as the query is already likely to retrieve a precise set of articles. Haynes et al. used internal validation, applying a split sample method to validate the filters [6]. External validation would be required, however, to conclusively establish the generalizability of Haynes et al.'s results [16]. While filters are traditionally designed and validated by hand-searching the medical literature, several recent studies focusing on the design or comparison of search filters have used relative recall of articles included in literature reviews [17–19]. This method regards articles identified in multiple information sources for reviews as representative of all the available evidence about a subject. Relative recall is then the proportion of these relevant articles that any specific system, filter, or tool retrieves. This method has been validated recently by Sampson et al. [20].

Systematic reviews conducted by researchers from the Cochrane organization are well suited to be the basis of a study of relative recall. They are aimed at retrieving all studies indexed in multiple databases or located by hand-searches that are relevant to clinical questions regarding therapy. This set of relevant articles is then critically appraised, resulting in a precise list of methodologically sound trials designed to answer a clinical question (Table 1). The Cochrane Database of Systematic Reviews [21] contains more than 3,000 reviews that contain links to references of included and excluded studies. Jenkins pointed out that filters need to be evaluated in a set of articles representative of the database it is designed for, reflect the type of questions it is used for, and preferably be tested for generalizability by external validation [22]. Haynes et al. only used a selection of randomized clinical trials from mainly clinically oriented journals, did not combine the filters with questions, or perform an external validation. Bachmann et al. studied the diagnostic Clinical Queries filters and suggested that their performance may be overrated [23].

As filters are designed to be used with user queries and precision depends on the prevalence of relevant articles, the real precision of the filter/query combination will depend on the precision of the query. Sensitivity and precision are inversely related in information retrieval [20], and previous epidemiological research has shown that sensitivity and specificity also vary with prevalence of a disease in ambiguous

Table 1
Comparison of the methods used to select studies concerning therapy by Haynes et. al. [7] and Cochrane reviews

	Haynes et al.	Cochrane
Study population	161 clinical journals	Multiple online databases and medical libraries
Identification of relevant articles	Hand-searching journals for: original study, review, general article, conference report, decision analysis, case report	Sensitive search strategy tailored to the question; hand-search based on references from relevant articles
Selection of relevant articles	Content pertained directly to therapy, prevention, or rehabilitation	Evaluation of abstracts for potential relevance to the question
Critical appraisal	Random allocation of participants to comparison groups, outcome assessment for at least 80% of those entering the investigation in 1 major analysis accounted for at least 1 follow-up assessment, and analysis consistent with study design	Extensive critical appraisal according to Cochrane manual; appraisal is required to adhere to the Cochrane standard

Table 2
Formulas to measure filter quality

	Methodologically strong and relevant	Not methodologically strong or irrelevant
Studies retrieved by filter	True positive (TP)	False positive (FP)
Studies filtered by filter	False negative (FN)	True negative (TN)

Sensitivity=TP/(TP+FN); specificity=TN/(FP+FN); precision=TP/(TP+FP).

tests [24]. Sensitivity and specificity of the filters can therefore be expected to vary with the precision of the query. This explains why external validation of filters in articles relevant to clinical questions is required. As the citations in Cochrane reviews are a selection of clinical trials that have been published in multiple information sources, including the whole of PubMed, and are related to a clinical question, the Cochrane Library is suitable as a source for external validation. The sensitivity, specificity, and precision found by Haynes et al. are representative of the use of the filter without a user query. The same parameters calculated from the Cochrane set are representative of the filter in combination with a query that has been extremely optimized. The sensitivity, specificity, and precision of the filters in daily queries will vary with every query but are likely to be between the values found in both sets. Combinations of filters and user query studies are likely to yield results somewhere between these extremes.

As all the reviews in the Cochrane database are related to treatment, the current study was limited to filters that are likely to improve search results for therapeutic studies. The primary goal was to determine the sensitivity, specificity, and precision of Clinical Queries filters in a set of articles relevant to a question concerning therapy. A secondary goal was to calculate the sensitivity, specificity, and precision of several other filters that are frequently advocated in the literature of evidence-based literature searching. Finally, the study sought to determine whether certain combinations of filters and limits are likely to improve search results. Articles included in and excluded from reviews available from the Cochrane database were used as the gold standard for calculating the sensitivity, specificity, and precision of several filters that are advocated in the evidence-based medicine literature.

METHODS

Selecting reviews

Background information from all systematic reviews published in the Cochrane Database of Systematic Reviews prior to January 2008 was retrieved. The database contains information not only about published reviews, but also information about reviews that have been withdrawn for an update or other reasons. Reviews that were withdrawn at the time of data collection were excluded from the study.

Extracting references

Screen-scraper Basic Edition [25] was used to extract literature information from the Cochrane Library and to retrieve article information from PubMed. This program extracts data from web pages based on the text and page structure. The web page containing review information was opened by screen-scraper, which then extracted references to studies included in and excluded from the review. Reviews that had no section listing included or excluded articles were excluded from the study. Most references in the reviews had a web link giving access to their abstracts in PubMed. This web link—which contains information about the journal title, author, volume, and page—was extracted by screen-scraper. References without links were excluded from analysis. The information provided was then sent to PubMed combined with the ''AND'' operator to retrieve the reference. When a unique reference to an article was obtained, the PubMed Unique Identifier (UID) was recorded.

Some references contained journal or author information that was not recognized by PubMed. This could be caused by the fact that Cochrane reviews use multiple sources that are often not available in PubMed (doctoral theses, journals not available in PubMed). These references were not relevant for our study and could safely be excluded. Misspelled references in the Cochrane database could, however, result in erroneous exclusion of references. To minimize the risk of erroneous exclusion of references, the authors broadened the query if no unique article was retrieved. The reference information was sent to PubMed using the journal title ''OR'' the author information. If a unique article identifier could still not be retrieved, the reference was excluded from the study. Reviews that did not provide one or more references that could be located in PubMed were also excluded. When references were repeated in the same reference section of a review, they were included only once. When the same reference was used in a different review, it was included in both reviews.

Selecting filters and limits

Clinical Queries filters and single term filters were selected from those in the original study by Haynes et al. [7]. Other PubMed filters (limits) were included because they have been described in the literature on evidence-based searching as helpful tools for narrowing a set of retrieved articles [26, 27]. As the limits used in our study and the Clinical Queries filters are

Table 3
Filters used in our study in Ovid and PubMed formats

	Ovid MEDLINE	PubMed
Sensitive Clinical Queries*	clinical trial.mp. OR clinical trial.pt. OR random:.mp. OR tu.xs	(clinical[Title/Abstract] AND trial[Title/Abstract]) OR clinical trials[MeSH Terms] OR clinical trial[Publication Type] OR random*[Title/Abstract] OR random allocation[MeSH Terms] OR therapeutic use[MeSH Subheading]
Specific Clinical Queries*	randomized controlled trial.pt. OR randomized controlled trial.mp	randomized controlled trial[Publication Type] OR (randomized[Title/Abstract] AND controlled[Title/Abstract] AND trial[Title/Abstract])
Single Term Medline Sensitive*	clinical trial.mp,pt	(clinical[Title/Abstract] AND trial[Title/Abstract]) OR clinical trials[MeSH Terms] OR clinical trial[Publication Type]
Single Term Medline Specific*	randomized controlled trial.pt	randomized Controlled Trial[ptyp]
Two Terms Medline Optimized*	randomized controlled trial.pt. OR randomized.mp. OR placebo.mp	randomized controlled trial[Publication Type] OR randomized[Title/Abstract] OR placebo[Title/Abstract]
Clinical trial	NA	Clinical Trial[ptyp]
English	NA	English[lang]
Humans	NA	''Humans''[MeSH Terms]
Abridged Index Medicus†	NA	jsubsetaim[text]

* Based on Haynes et al. [7].
† The Abridged Index Medicus is a set of core clinical journals. A complete list of journals is available at: http://www.nlm.nih.gov/bsd/aim.html.
mp=multiple posting (term appears in title, abstract, or MeSH heading); pt=publication type; tu=therapeutic use subheading; xs=exploded subheading; :=truncation.

both filters by definition and PubMed handles them comparably, they will all be referred to here as filters. Full descriptions of the filters used in this study and their relationship to those used by Haynes et al. [7] can be found in Table 3.

Applying filters

For each review, both included and excluded references were sent to PubMed separately with and without filters. Articles included in each Cochrane review qualified as true positives if they were not filtered by the filter and as false negatives if they were. Articles excluded in each review qualified as false positives if they were not filtered by the filter and as true negatives if they were. From these results, sensitivity, specificity, and precision were calculated for each review (Table 2).

Combining filters

Combining filters with an ''AND'' potentially increases precision. However, combinations of filters are only effective if they use different filter criteria, because the use of similar filters will yield roughly the same articles. The magnitude of overlap in articles retrieved by different filters is difficult to predict if the selection criteria do not match. Because the Clinical Queries, Clinical Trial, Two Terms Medline Optimized, and Single Term Medline filters use roughly the same filter criteria (Table 3) and combining them would only result in small changes in precision, these filters can be regarded as a group. The Humans, English, and Abridged Index Medicus (AIM) filters are distinct. The first two are based on MeSH terms, and the last is based on a set of core clinical journals. The most specific filter from the four classes was selected first and then combined with all other filters in turn to determine whether combinations of two filters can increase precision.

The four filters with highest specificity regardless of possible overlap in the use of terms were then combined to determine whether combining several precise filters, although they have some terms in common, could be used to increase the precision of a search. Finally, the Humans and English filters were added to this precise combination. Sensitivity, specificity, and precision measures were calculated for all the combinations of filters for each review.

Analyzing statistics

The values for sensitivity, specificity, and precision of all reviews were combined, and mean values were calculated. SPSS version 14.0 was used to determine the standard error of the mean for these measures. The tables show the 95% confidence intervals based on the standard error of mean.

RESULTS

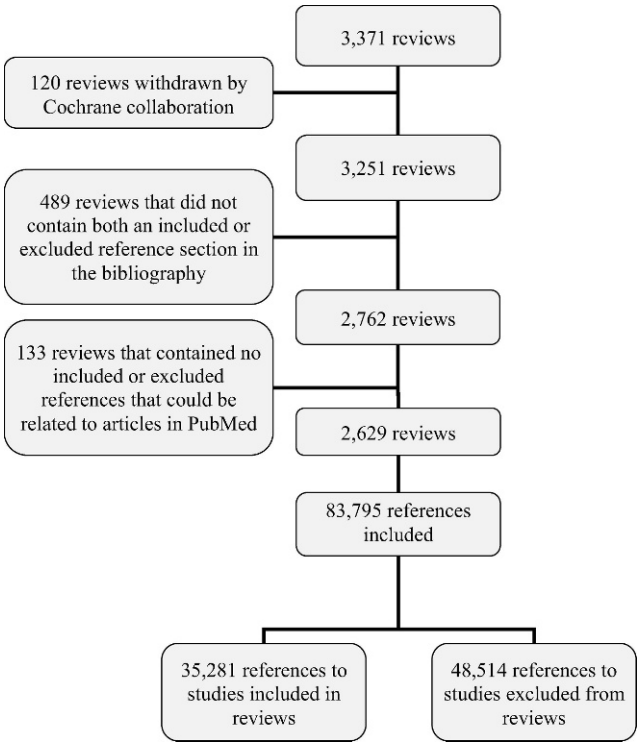
Two thousand six hundred twenty-nine of the 3,371 reviews available in the Cochrane database contained both included and excluded articles that could be retrieved from PubMed. These reviews yielded 83,975 references to articles (Figure 1), an average of 13.4 included and 18.5 excluded references per review.

Sensitivity, specificity, and precision for each review were determined, and the means were calculated (Table 4). The highest sensitivity was achieved using a Sensitive Clinical Queries, English, or Humans filter, at the cost of very low specificity. The highest specificity was achieved using the AIM filter.

Combinations leading to optimal precision

Results from combining the highest-specificity AIM filter as the basic filter with other filters are shown in Table 5. The highest precision was achieved using the

Figure 1
Overview of reviews from the Cochrane Database



Single Term Medline Specific, Two Terms Medline Optimized, or Specific Clinical Queries filter in combination with AIM.

Combining the top four specific filters—AIM, Specific Clinical Queries, Single Term Medline Specific, and Two Terms Medline Optimized—did not increase specificity or precision: sensitivity, 30.2%; specificity, 85.9%; and precision, 65.2%. The English and Humans filters did not improve the results: adding both these filters to the optimal combination of the AIM and Single Term Medline Specific filters resulted in a sensitivity of 26.8%, specificity of 86.7%, and precision of 59.5%.

DISCUSSION

This study shows the sensitivity, specificity, and precision for Clinical Queries and selected other PubMed filters using studies included in and excluded from Cochrane reviews as the gold standard. The sensitivities and specificities found in this study were much lower than those reported by Haynes et al., but the precision was higher [7]. There are several reasons for these differences.

1. Cochrane database and relative recall

Relative recall is the proportion of articles found in one system in relation to the information retrieved from several databases. Relative recall for validation of filters has been proposed as an alternative to hand-searching [20]. The Cochrane reviews use many sources to find information. In this study, relative recall was calculated using a set of articles retrieved from the whole of PubMed. The filters developed by Haynes et al. were constructed using a set of articles retrieved from a set of core clinical journals [7]. Part of the difference in sensitivity, specificity, and precision may be related to the difference in the pool of articles used.

2. Split sample validation

Another reason why the values here were different from those of Haynes et al. may be that Haynes validated the filters internally using a split sample technique [7]. External validation is often preferred because internal validation can result in overfitting, leading to overestimation of sensitivity and specificity [16, 28, 29].

3. Critical appraisal

The articles selected for Cochrane reviews are evaluated for methodological rigor. If they are not methodologically sound or do not provide the data necessary to adequately answer a question about therapy, they are excluded from the review. Haynes et al. used less rigorous criteria than the Cochrane reviewers to establish methodological quality [7].

Table 4
Mean sensitivity, specificity, and precision of filters retrieving articles included and excluded for systematic reviews in the Cochrane Database

PubMed filters*	Sensitivity		Specificity		Precision	
	mean %	(95% CI)	mean %	(95% CI)	mean %	(95% CI)
Sensitive Clinical Queries	92.7	(92.1–93.3)	16.1	(15.1–17.1)	49.5	(48.5–50.5)
Specific Clinical Queries	78.2	(77.2–79.2)	52.0	(50.8–53.2)	60.4	(59.4–61.4)
Single Term Medline Sensitive	88.1	(87.3–88.9)	32.6	(31.4–33.8)	54.1	(53.1–55.1)
Single Term Medline Specific	78.0	(77.0–79.0)	52.3	(51.1–53.5)	60.6	(59.6–61.6)
Two Terms Medline Optimized	82.0	(81.0–83.0)	46.9	(45.7–48.1)	59.0	(58.0–60.0)
Clinical trial	87.3	(86.5–88.1)	34.8	(33.6–36.0)	54.7	(53.7–55.7)
English	95.9	(95.5–96.3)	8.9	(8.3–9.5)	48.2	(47.2–49.2)
Humans	97.9	(97.5–98.3)	3.3	(2.9–3.7)	47.2	(46.2–48.2)
Abridged Index Medicus	35.9	(34.7–37.1)	74.2	(73.2–75.2)	53.8	(52.4–55.2)

* PubMed filters described in Table 3.

Table 5
Combination of the Abridged Index Medicus (AIM) filter with other filters

Filter combined with AIM*	Sensitivity		Specificity		Precision	
	mean %	(95% CI)	mean %	(95% CI)	mean %	(95% CI)
Sensitive Clinical Queries	34.3	(33.1–35.5)	77.5	(76.5–78.5)	56.5	(55.1–57.9)
Specific Clinical Queries	30.3	(29.1–31.5)	85.8	(85.0–86.6)	65.0	(63.6–66.4)
Single Term Medline Sensitive	33.1	(31.9–34.3)	81.5	(80.7–82.3)	60.6	(59.2–62.0)
Single Term Medline Specific	30.3	(29.3–31.3)	85.9	(85.1–86.7)	65.2	(63.8–66.6)
Two Terms Medline Optimized	31.4	(30.2–32.6)	84.7	(83.9–85.5)	64.2	(62.8–65.6)
Clinical trial	32.8	(31.6–34.0)	82.0	(81.2–82.8)	61.3	(59.9–62.7)
English	35.9	(34.7–37.1)	74.2	(73.2–75.2)	53.8	(52.4–55.2)
Humans	35.8	(34.6–37.0)	72.4	(71.4–73.4)	53.9	(52.5–55.3)

* PubMed filters described in Table 3.

Some of the differences between the results in the two studies may therefore be explained by a difference in critical appraisal. As the articles in the current study are all relevant to a medical topic, the lower sensitivity and specificity reflect the fact that the Clinical Queries filters are not very effective in selecting only truly methodologically sound trials.

4. Set of articles reflecting precise clinical query

The search methodology that the Cochrane Collaboration employs to identify articles is extremely sensitive, attempting to retrieve all available knowledge about a topic. The large set of articles retrieved is then evaluated by researchers to identify those that are potentially relevant for the review, resulting in a set of articles that are very precisely related to a clinical question about therapy. The refined set of articles therefore represents a very precise query. Haynes et al. did not design and validate their filters in sets of articles retrieved on a particular topic [7]. Their filters were designed to be used in combination with user queries. It is unlikely that physicians, while searching on the spot, can articulate queries so precisely that the resulting article sets are comparable to those selected for the Cochrane reviews. The real sensitivity, specificity, and precision of Clinical Queries, when used in actual practice, are likely to be somewhere between the results reported by Haynes et al. [7] and those reported here.

5. Combinations of filters

The combination of the specific Single Term Medline, Two Terms Medline Optimized, or Specific Clinical Queries filters with the AIM filter yields highest specificity and precision (65%). Whether physicians should be discouraged from using sources other than the few clinical journals included in AIM is open to debate, but limiting the search to those journals can be justified if too many results are retrieved by a query. Adding other filters could not increase precision in our set of references. The Humans and English filters have low specificities because they are not related to the methodological quality of a study. Our results show that when they are used with a precise set of filters, they result in lower precision. The Humans filter may help in a topic that is likely to retrieve a considerable number of animal studies, but use of these filters should not be taught for evidence-based searching.

6. Implications for practice

The results of this study show that it is possible to reach moderate precision using specific filters in precise article sets related to a clinical question. Although the study by Haynes et al. suggests that specific filters are able to identify methodologically sound studies about therapy, the Clinical Queries filters from that study mainly select randomized clinical trials. As randomization is only one of the criteria that determine the scientific quality of a study and many clinical questions cannot be studied with randomized controlled trials, the suggestion that the filter will help retrieve only methodologically sound studies may result in inappropriate use. Because most articles in this study's dataset are designed as randomized controlled trials, in contrast to studies about other subjects, it is no surprise that ''randomized Controlled Trial[ptp]'' was confirmed as the most specific filter term.

As the precision of the query, and thus the need for a filter, depends on searchers' skills, skilled searchers are less likely to improve search results using methodological filters than inexperienced searchers using less precise queries. If physicians are taught to create very accurate queries—adequately describing the patient, intervention, possible alternative interventions and outcomes of interest, and study types they are interested in—the precision of the queries is likely to be better than can be reached by combining queries with Clinical Queries filters. The advantage of adequately formed queries over the use of Clinical Queries filters is that physicians understand what they are searching for, can search for more study types than trials alone, and will retrieve articles about subjects that cannot be studied with randomized controlled trials.

Sensitivities, specificities, and precisions varied widely, because many reviews contained references to a small set of relevant articles. Questions raised during medical practice are likely to show the same variation in relevant information sources. Physicians should also expect a wide array of values for these filters in daily practice.

Sensitive Clinical Queries filters are more suitable for reviews of the literature than patient-related searches as they will retrieve many irrelevant articles. In reviews of the literature, it is crucial that all the relevant literature is retrieved by the filter. As the sensitive filters seem to filter a significant proportion of relevant articles, they may not be appropriate for conducting reviews.

AIM is a very specific filter that can be used in combination with other filters, but the Humans and English filters do not discriminate between methodologically sound and unsound studies and should be used with care. Contrary to advice given in evidence-based literature, simply combining several filters is not a sensible way to increase the precision of a search.

Limitations

Although the set of articles in this study was derived from the whole of PubMed, it still consisted of a set of articles selected for a purpose that is not entirely comparable to the daily medical care situation. Only original studies are included as references in the Cochrane Library, with preference for randomized clinical trials. References to other potential evidence-based sources of information, such as reviews, are not included in the dataset.

Filters designed for diagnostic or prognostic studies or other subjects relevant to clinical problem solving could not be tested, because Cochrane reviews addressing these topics do not contain a reference list that is suitable for automatic retrieval.

A considerable number of reviews were not included because no references to articles in an included or excluded section could be retrieved from PubMed. Occasional typing errors were excluded by searching with the author ''OR'' the journal information. Only if both were spelled incorrectly, could the reference not be traced. Because misspelling is just as likely to occur in relevant as in irrelevant articles, this is not a source of bias.

The search filters were not combined with user-generated queries. Because user-generated queries are likely to show wide variability, several queries would be required per review. That would considerably limit the number of reviews that could be included. Articles that are potentially relevant for Cochrane reviews are usually studies conducted in human subjects, and the publication language is usually English. The sensitivity and specificity of the Humans and English filters are likely to be very different in combination with user queries. We included these filters to emphasize that they do not identify clinically sound studies, and their use should not be recommended for evidence-based searching.

CONCLUSIONS

Clinical Queries filters have been designed to help physicians limit search results, achieving high sensitivity or specificity, depending on the filter, but limited precision. These filters have been previously validated by internal validation on a subset of clinical journals, not related to clinical questions, limiting the generalizability of the results. This study is the first to perform an external validation of the Clinical Queries filters for therapy using references in Cochrane reviews as a gold standard. Sensitivity and specificity were low in a set of articles relevant to clinical questions. The highest precision was achieved by the Single Term Medline Specific filter, which uses ''randomized Controlled Trial[ptyp]'' as the filter criterion. Combining the AIM filter with the Specific Clinical Queries, Two Terms Medline Optimized, or the Single Term Medline Specific filters resulted in a slightly higher precision. The results indicated that PubMed filters can help reduce the number of articles to be read but are unlikely to compensate for ill-formed queries. Moreover, Specific Clinical Queries filters will filter the best evidence available in questions that cannot be answered with randomized trials and retrieve randomized controlled trials that are not clinically sound according to CONSORT criteria. It is therefore likely that adequate formulation of queries, reflecting the question and the patient, aimed at retrieving articles that are likely to report the required information, will yield better search results than the use and combination of search filters. Further research is needed to determine the additional value of filters in adequately formed queries.

ROLE OF THE FUNDING SOURCE

This study was funded by the *Vereniging Nederlands Tijdschrift voor Geneeskunde (Association Dutch Journal of Medicine)*.

REFERENCES

1. US National Library of Medicine. PubMed [Internet]. Bethesda, MD: The Library; 2008 [cited 19 Dec 2008]. <http://www.ncbi.nlm.nih.gov/sites/entrez/>.
2. Hoogendam A, Stalenhoef AF, Robbe PF, Overbeke AJ. Answers to questions posed during daily patient care are more likely to be answered by UpToDate than PubMed. J Med Internet Res. 2008;10(4):e29. DOI: 10.2196/jmir.1012.
3. Guyatt GH, Rennie D, eds. Users' guides to the medical literature: a manual for evidence-based clinical practice. Chicago, IL: American Medical Association Press; 2002.
4. Straus SE, Richardson WS, Glasziou P, Haynes RB. Evidence-based medicine: how to practice and teach EBM. 3rd ed. Edinburgh, Scotland, UK: Churchill Livingstone; 2005.
5. The CONSORT Group. The CONSORT statement [Internet]. The Group; 2008 [cited 19 Dec 2008]. <http://www.consort-statement.org>.
6. Haynes RB, Wilczynski NL. Optimal search strategies for retrieving scientifically strong studies of diagnosis from Medline: analytical survey. BMJ. 2004 May 1;328(7447):1040. DOI: 10.1136/bmj.38068.557998.EE.
7. 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. 2005 May 21;330(7501):1162–3. DOI: 10.1136/bmj.38446.498542.8F.

8. Montori VM, Wilczynski NL, Morgan D, Haynes RB. Optimal search strategies for retrieving systematic reviews from Medline: analytical survey. BMJ. 2005 Jan 8;330(7482):68. DOI: 10.1136/bmj.38336.804167.47.
9. Wilczynski NL, Haynes RB, Lavis JN, Ramkissoonsingh R, Arnold-Oatley AE. Optimal search strategies for detecting health services research studies in MEDLINE. CMAJ. 2004 Nov 9;171(10):1179–85.
10. Wilczynski NL, Haynes RB. Developing optimal search strategies for detecting clinically sound prognostic studies in MEDLINE: an analytic survey. BMC Med. 2004 Jun 9;2:23. DOI: 10.1186/1741-7015-2-23.
11. Wong SS, Wilczynski NL, Haynes RB. Developing optimal search strategies for detecting clinically relevant qualitative studies in MEDLINE. Stud Health Technol Inform. 2004;107(pt 1):311–6.
12. Ely JW, Osheroff JA, Ebell MH, Bergus GR, Levy BT, Chambliss ML, Evans ER. Analysis of questions asked by family doctors regarding patient care. BMJ. 1999 Aug 7;319(7206):358–61.
13. Gorman PN, Ash J, Wykoff L. Can primary care physicians' questions be answered using the medical journal literature? Bull Med Libr Assoc. 1994 Apr;82(2): 140–6.
14. Schwartz K, Northrup J, Israel N, Crowell K, Lauder N, Neale AV. Use of on-line evidence-based resources at the point of care. Fam Med. 2003 Apr;35(4):261–3.
15. Westbrook JI, Coiera EW, Gosling AS. Do online information retrieval systems help experienced clinicians answer clinical questions? J Am Med Inform Assoc. 2005 May–Jun;12(3):315–21. DOI: 10.1197/jamia.M1717.
16. Bleeker SE, Moll HA, Steyerberg EW, Donders AR, Derksen-Lubsen G, Grobbee DE, Moons KG. External validation is necessary in prediction research: a clinical example. J Clin Epidemiol. 2003 Sep;56(9):826–32. DOI: 10.1016/S0895-4356(03)00207-5.
17. 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 May;58(5):444–9. DOI: 10.1016/j.jclinepi.2004.09.011.
18. 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 Feb;31(1):150–3.
19. 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 Sep;20(3):150–9. DOI: 10.1046/j.1365-2532.2003.00427.x.
20. Sampson M, Zhang L, Morrison A, Barrowman NJ, Clifford TJ, Platt RW, Klassen TP, Moher D. An alternative to the hand searching gold standard: validating methodological search filters using relative recall. BMC Med Res Methodol. 2006 Jul 18;6:33. DOI: 10.1186/1471-2288-6-33.
21. The Cochrane Collaboration. The Cochrane library [Internet]. Oxford, UK: The Collaboration; c2004–2006 [2008; cited 19 Dec 2008]. <http://www.thecochranelibrary.com>.
22. Jenkins M. Evaluation of methodological search filters—a review. Health Info Libr J. 2004 Sep;21(3):148–63. DOI: 10.1111/j.1471-1842.2004.00511.x.
23. 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 Nov;9(6):653–8. DOI: 10.1197/jamia.M1124.
24. Brenner H, Gefeller O. Variation of sensitivity, specificity, likelihood ratios and predictive values with disease prevalence. Stat Med. 1997 May 15;16(9):981–91.
25. Screen-scraper Basic Edition. Version 4.0 [computer program]. e-kiwi; 21 Jan 2008.
26. Greenhalgh T. How to read a paper. the Medline database. BMJ. 1997 Jul 19;315(7101):180–3.
27. Ebbert JO, Dupras DM, Erwin PJ. Searching the medical literature using PubMed: a tutorial. Mayo Clin Proc. 2003 Jan;78(1):87–91.
28. Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. Ann Intern Med. 1999 Mar 16;130(6):515–24.
29. Van den Bruel A, Aertgeerts B, Buntinx F. Results of diagnostic accuracy studies are not always validated. J Clin Epidemiol. 2006 Jun;59(6):559–66.

AUTHORS' AFFILIATIONS

Arjen Hoogendam, MD, A.Hoogendam@aig.umcn.nl, Specialist in Internal Medicine, Department of General Internal Medicine and Department of Medical Informatics; **Pieter F. de Vries Robbé, MD, PhD,** P.deVriesRobbe@mi.umcn.nl, Professor, Department of Medical Informatics; **Anton F. H. Stalenhoef, MD, PhD, FRCP,** A.Stalenhoef@AIG.umcn.nl, Professor, Department of General Internal Medicine; **A. John P. M. Overbeke, MD, PhD,** J.Overbeke@mi.umcn.nl, Professor, Department of Medical Informatics; Radboud University Nijmegen Medical Centre, P.O. Box 9101, Nijmegen, The Netherlands

Received October 2008; accepted February 2009