MEDLINE clinical queries are robust when searching in recent publishing years

Nancy L Wilczynski,¹ K Ann McKibbin,¹ Stephen D Walter,² Amit X Garg,³
R Brian Haynes^{1,4}

¹Health Information Research Unit, Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario, Canada

²Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario, Canada

³Department of Medicine, Department of Epidemiology and Biostatistics, Western University, London, Ontario, Canada

⁴Hamilton Health Sciences, Hamilton, Ontario, Canada

Correspondence to

Dr Nancy L Wilczynski, Health Information Research Unit, Department of Clinical Epidemiology and Biostatistics, McMaster University, CRL 125, 1280 Main Street West, Hamilton, ON L8S 4K1, Canada; wilczynn@mcmaster.ca.

Received 30 April 2012

Accepted 2 September 2012

Published Online First

27 September 2012

ABSTRACT

Objective To determine if the PubMed and Ovid MEDLINE clinical queries (which were developed in the publishing year 2000, for the purpose categories therapy, diagnosis, prognosis, etiology, and clinical prediction guides) perform as well when searching in current publishing years.

Methods A gold standard database of recently published research literature was created using the McMaster health knowledge refinery (http://hiru.mcmaster.ca/hiru/HIRU_McMaster_HKR.aspx) and its continuously updated database, McMaster PLUS (http://hiru.mcmaster.ca/hiru/HIRU_McMaster_PLUS_projects.aspx). This database contains articles from over 120 clinical journals that are tagged for meeting or not meeting criteria for scientific merit and clinical relevance. The clinical queries sensitive ('broad') and specific ('narrow') search filters were tested in this gold standard database, and sensitivity and specificity were calculated and compared with those originally reported for the clinical queries.

Results In all cases, the sensitivity of the highly sensitive search filters and the specificity of the highly specific search filters did not differ substantively when comparing results derived in 2000 with those derived in a more current database. In addition, in all cases, the specificities for the highly sensitive search filters and the sensitivities for the highly specific search filters remained above 50% when testing them in the current database.

Discussion These results are reassuring for modern-day searchers. The clinical queries that were derived in the year 2000 perform equally well a decade later.

Conclusion The PubMed and Ovid MEDLINE clinical queries have been revalidated and remain a useful public resource for searching the world's medical literature for research that is most relevant to clinical care.

INTRODUCTION

Scientific knowledge grows quickly and is often first accessible in peer review publications via online databases such as PubMed and Ovid MEDLINE. MEDLINE currently contains over 21 million references to journal articles and adds between 10 000 and 20 000 references each week.¹ Clinicians,²⁻³ patients,⁴ and the public⁵ often search the internet to find answers to clinical questions. For example, 1.8 billion searches were conducted in PubMed in 2011.⁶ However, finding the current best evidence to answer clinical questions is a challenge because strong clinical studies are a very small fraction of the MEDLINE database. To help alleviate this challenge, clinical queries were designed for use in PubMed and Ovid MEDLINE

to aid in retrieving the current best evidence for healthcare.

Clinical queries are online search filters designed to improve the retrieval of scientifically strong and clinically relevant articles from the US National Library of Medicine's (NLM) PubMed database (<http://www.ncbi.nlm.nih.gov/pubmed/clinical>) and Ovid MEDLINE. These search filters were developed and validated using data from the publishing year 2000. The clinical queries have been available in PubMed since that time. A recent study showed that MEDLINE is a valuable resource for just-in-time answers to clinical questions and that using PubMed clinical queries facilitates the timely retrieval of results.⁷ In addition, a study published in 2012 showed that the use of the PubMed clinical queries improved the efficiency of physician searches by improving the ratio of relevant to non-relevant articles.⁸ The current typical usage of the clinical queries in PubMed is 10 000 daily page views from 3300 users (National Center for Biotechnology Information, User Services, 8 March 2012, personal communication).

More than a decade has passed since the clinical query filters were developed. Staff at the NLM routinely review and revise search filters that have been developed at NLM to account for new medical subject headings (MeSH), vocabulary, etc., and expressed concerns, on behalf of the users of the clinical queries, that these search filters could also benefit from analysis and review (Ione Auston, National Information Center on Health Services Research and Health Care Technology, personal communication). It is possible that the performance of the clinical query filters may have declined over time. Declining performance may be due to changes in indexing rules and/or the practice of applying the index terms. Declining performance could also be due to changes in the reporting of the methods in the articles. Authors are continually encouraged to be more transparent and consistent in their reporting, as outlined in reporting standards such as CONSORT,⁹ PRISMA,¹⁰ and STARD.¹¹ A change in the reporting of research methods could have made our choice of methodological text words and indexing terms less appropriate over time. In addition, declining performance may be due to changes in the search engines used by PubMed and Ovid and in their conventions for handling indexing terms and text words. These concerns prompted the following questions: Are the clinical queries (which were developed in the publishing year 2000, for the purpose categories therapy, diagnosis, prognosis, etiology, and clinical prediction guides) robust

when searching in current publishing years? If the search filters do not perform as well, can they be re-calibrated to optimize performance?

METHODS

To determine the performance of the clinical queries when searching in more current publishing years, we used the McMaster health knowledge refinery¹² and its continuously updated database, McMaster PLUS.¹³ The McMaster PLUS database is created by research staff in the Health Information Research Unit at McMaster University who currently critically appraise the content of more than 120 clinically relevant journals¹⁴ on an ongoing basis. Articles that meet explicit critical appraisal criteria¹⁵ are included in the database after they have been rated for clinical relevance and newsworthiness by a worldwide panel of practising clinicians. For this study, the McMaster PLUS database, which contains articles that meet criteria for scientific merit and clinical relevance in journals indexed by MEDLINE, was augmented with all articles, from the same journals, that do not meet explicit criteria for scientific merit and clinical relevance. This augmented database thus constituted the modern equivalent of a hand searched, gold standard database.

In this newly constructed database, we tested the performance of the broad (highly sensitive) and narrow (highly specific) clinical query search filters for therapy,¹⁶ diagnosis,¹⁷ prognosis,¹⁸ etiology,¹⁹ and clinical prediction guides.²⁰ The categorization and methods used for article inclusion in the McMaster PLUS database are almost identical to those used in 2000 when originally developing the clinical queries. Table 1 shows the definitions used for the categorizations of purpose (eg, therapy) in 2000 when constructing the clinical hedges database, and in the years 2003–11 when constructing the McMaster PLUS database. As noted in table 1, the definitions for diagnosis, prognosis and clinical prediction guides are identical. The purpose category therapy is slightly different. In the McMaster PLUS database studies testing therapy and quality improvement interventions are two separate ‘purpose’ categories, whereas they are combined in the clinical hedges database. Therefore, when testing the highly sensitive and highly specific therapy filters in the newly constructed database, the categories

of therapy and quality improvement were combined. In addition, a difference in the definitions was applied for the purpose category etiology when comparing the clinical hedges database and the McMaster PLUS database.

Table 2 shows the methodological criteria used to determine if an article was methodologically sound in 2000 when constructing the clinical hedges database and in the years 2003–11 when constructing the McMaster PLUS database. The criteria used for diagnosis, prognosis and clinical prediction guides are identical. For therapy, all criteria are identical except for the additional criterion applied by the research staff for the McMaster PLUS database added on 2 March 2011: ‘outcome measure of known or probable clinical importance’. For etiology, the change in definition noted in table 1 led to the change in the methodological criteria applied, as noted in table 2.

The sample size for this study was based on the article by Yao and colleagues,²¹ which provides guidance for determining sample size when developing search filters. According to Yao and colleagues,²¹ at least 99 methodologically sound articles on therapy are needed to test search filters with reasonable precision (95% CI $\pm 5\%$). The corresponding number of articles for diagnosis and prognosis are 352 and 384, respectively. Sample size requirements for other purpose categories (eg, etiology) have not been formally determined.

As in our previous research, the analysis was conducted using a diagnostic testing model with sensitivity and specificity being calculated. The augmented McMaster PLUS database served as the gold standard and combinations of methodological search terms (MeSH and text words) constituted the test. We tested the clinical query search filters that are currently in use in PubMed and Ovid MEDLINE to determine if they are robust when searching in current publishing years. Programmers working within the Health Information Research Unit modified the programming code and the clinical hedges website interface to accommodate this testing. The Pearson χ^2 test for two proportions was used to determine if there was a statistically significant difference between the results (eg, sensitivity) when comparing results derived in the 2000 clinical hedges database with those derived in a more current database; a two-tailed p value of less than 0.05 indicated a statistically significant difference.

Table 1 Purpose category definitions used when constructing the clinical hedges and McMaster PLUS databases (bold, italicized text denotes differences in definitions)

Purpose category	Definitions used in the clinical hedges database (2000)	Definitions used in the McMaster PLUS database (2003–11)
Therapy	Content pertains directly to an intervention for therapy (including adverse effects studies), prevention, rehabilitation, quality improvement, or continuing medical education.	Content pertains directly to an intervention for therapy (including adverse effects studies), prevention, or rehabilitation. <i>Quality improvement and continuing medical education are a separate category (with same methodological criteria as therapy).</i>
Diagnosis	Content pertains directly to using a tool to arrive at a diagnosis of a disease or condition.	Content pertains directly to using a tool to arrive at a diagnosis of a disease or condition.
Prognosis	Content pertains directly to the prediction of the clinical course or the natural history of a disease or condition with the disease or condition existing at the beginning of the study.	Content pertains directly to the prediction of the clinical course or the natural history of a disease or condition with the disease or condition existing at the beginning of the study.
Etiology	Content pertains directly to determining if there is an association between an exposure and a disease or condition. The question is ‘What causes people to get a disease or condition?’	<i>Before 4 August 2009 same definition as in 2000 applied. After August 2004 definition changed to: content pertains directly to assessing the adverse effects/harm/safety of drugs prescribed by a physician. Design other than an RCT.</i>
Clinical prediction guides	Content pertains directly to the prediction of some aspect of a disease or condition. Guide deals with treatment, diagnosis, prognosis, or etiology. Guide must include ≥ 2 factors. Patients followed up over time.	Content pertains directly to the prediction of some aspect of a disease or condition. Guide deals with treatment, diagnosis, prognosis, or etiology. Guide must include ≥ 2 factors. Patients followed up over time.

RCT, randomized controlled trial.

Table 2 Methodological rigor criteria used when constructing the clinical hedges and the McMaster PLUS databases (bold, italicized text denotes differences in criteria)

Purpose category	Methodological criteria used in the clinical hedges database (2000)	Methodological criteria used in the McMaster PLUS database (2003–11)
Therapy	Random allocation of participants to comparison groups; Outcome assessment of at least 80% of those entering the investigation accounted for in one major analysis at any given follow up assessment; Analysis consistent with study design.	Random allocation of participants to comparison groups; Outcome assessment of at least 80% of those entering the investigation accounted for in one major analysis at any given follow up assessment; Analysis consistent with study design. <i>Before 2 March 2011 same criteria as in 2000 were applied. After 2 March 2011—outcome measure of known or probable clinical importance are applied by research staff.</i>
Diagnosis	Inclusion of a spectrum of participants; Objective diagnostic ('gold') standard or current clinical standard for diagnosis; Participants received both the new test and some form of the diagnostic standard; Interpretation of diagnostic standard without knowledge of test result and vice versa; Analysis consistent with study design.	Inclusion of a spectrum of participants; Objective diagnostic ('gold') standard or current clinical standard for diagnosis; Participant received both the new test and some form of the diagnostic standard; Interpretation of diagnostic standard without knowledge of test result and vice versa; Analysis consistent with study design.
Prognosis	Inception cohort of individuals all initially free of the outcome of interest; Follow-up of $\geq 80\%$ of patients until the occurrence of a major study end point or to the end of the study; Analysis consistent with study design.	Inception cohort of individuals all initially free of the outcome of interest; Follow-up of $\geq 80\%$ of patients until the occurrence of a major study endpoint or to the end of the study; Analysis consistent with study design.
Etiology	Observations concerned with the relationship between exposures and putative clinical outcomes; Data collection is prospective; Clearly identified comparison group(s);* Blinding of observers of outcome to exposure.	<i>Before 4 August 2009 same criteria as in 2000 applied. After 4 August 2009 criteria changed to—data collection is prospective; Clearly identified comparison group(s)* (does not include RCT); Blinding of observers of outcomes to exposures; Must have at least 80% follow-up.</i>
Clinical Prediction Guides	Guide is generated in one or more sets of real patients (training set); Guide is validated in another set of real patients (test set).	Guide is generated in one or more sets of real patients (training set); Guide is validated in another set of real patients (test set).

*Randomized controlled trials (RCT), quasi-randomized controlled trials, non-randomized controlled trials, cohort studies with case-by-case matching or statistical adjustment to create comparable groups, or nested case-control studies.

Constructing the gold standard database

A list of journal titles that have been critically appraised to produce content for the McMaster PLUS database as well as other services and products (eg, *ACP Journal Club*) is maintained within the Health Information Research Unit and shows that 389 journal titles have been reviewed over the years 1997–2011. Journal titles are added and removed from the list based on their performance (ie, whether they publish clinical content that is of high methodological quality). Because the McMaster PLUS database contains articles published from 2003 on, journal titles that had been dropped from the critical appraisal process before 2003 were removed from the list of potential journal titles that could be included in the database for this study. The revised list of journal titles that could be potentially included in the gold standard database for this study was 184 titles (available from the authors on request).

Therapy search filters

The test for therapy search filters was conducted in a gold standard database that contained articles published in 2010. The rationale for compiling a 2010 gold standard database was that the critical appraisal criteria for therapy articles changed on 2 March 2011, as noted above and in table 2. The journal title list for testing the therapy filters included those journals that were read for the entire year of 2010 and those that did not have a name change during 2010. Journals that were not indexed in MEDLINE were also deleted from the list. As a result, 114 journal titles were contained in the 2010 database for testing the therapy search filters. The clinical queries tested for therapy are shown in table 3.

Etiology search filters

The test for etiology search filters was conducted in a database that contained articles published in the years 2007 and 2008.

The rationale for compiling a 2007–8 gold standard database was that the methodological criteria for etiology changed on 4 August 2009, as noted above and in table 2. The journal list for testing the etiology filters included those journals that were read for the entire time period of 2007 and 2008, and did not have a name change during that time period. Journals that were not indexed in MEDLINE were deleted from the list. As a result, 136 journal titles were contained in the 2007–8 database for testing the etiology search filters. The clinical queries tested for etiology are shown in table 3.

Diagnosis, prognosis and clinical prediction guide search filters

The test for the diagnosis, prognosis and clinical prediction guide search filters was conducted in a database that contained articles published in the years 2003–10. The rationale for compiling a 2003–10 gold standard database was that this time period would be required to obtain a sufficient sample size of methodologically sound articles in each purpose category. The journal list for testing these filters included those journals that were read for the entire time period of 2003–10, and did not have a name change during that time. Journals that were not indexed in MEDLINE were deleted from the list. As a result, 78 journal titles were contained in the 2003–10 database for testing the diagnosis, prognosis, and clinical prediction guide search filters. The clinical queries tested for diagnosis, prognosis, and clinical prediction guides are shown in table 3.

RESULTS

The breakdown of the number of citations included in the clinical hedges database and the McMaster PLUS databases by purpose category is shown in table 4.

The results for the sensitive therapy, diagnosis, prognosis, etiology and clinical prediction guide search filters are shown in

Table 3 Search filters that were tested

Purpose category	Sensitive search filter Ovid syntax PubMed syntax	Specific search filter Ovid syntax PubMed syntax
Therapy	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))	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)))
Diagnosis	sensitiv.:mp. OR diagnos.:mp. OR di.fs. (sensitiv*(Title/Abstract) OR sensitivity and specificity (MeSH Terms) OR diagnos*(Title/Abstract) OR diagnosis(MeSH:noexp) OR diagnostic * (MeSH:noexp) OR diagnosis,differential(MeSH: noexp) OR diagnosis(Subheading:noexp))	specificity.tw. (specificity(Title/Abstract))
Prognosis	incidence.sh. OR exp mortality OR follow-up studies.sh. OR prognos.:tw. OR predict.:tw. OR course.:tw. (incidence (MeSH:noexp) OR mortality(MeSH Terms) OR follow up studies(MeSH:noexp) OR prognos*(Text Word) OR predict*(Text Word) OR course*(Text Word))	prognos.:tw. OR first episode.tw. OR cohort.tw. (prognos*(Title/Abstract) OR (first(Title/Abstract) AND episode(Title/Abstract)) OR cohort(Title/Abstract))
Etiology	risk:mp. OR exp cohort studies OR between group.:tw. (risk* (Title/Abstract) OR risk*(MeSH:noexp) OR risk *(MeSH:noexp) OR cohort studies(MeSH Terms) OR group*(Text Word))	relative risk:tw. OR risks.tw. OR cohort stud.:mp. ((relative (Title/Abstract) AND risk* (Title/Abstract)) OR (relative risk(Text Word)) OR risks(Text Word) OR cohort studies(MeSH: noexp) OR (cohort(Title/Abstract) AND stud*(Title/Abstract)))
Clinical prediction guides	predict:mp. OR scor.:tw. OR observ:mp. (predict*(Title/Abstract) OR predictive value of tests (MeSH Terms) OR scor*(tiab) OR observ*(Title/ Abstract) OR observer variation(MeSH Terms))	validation.tw. OR validate.tw. (validation(Title/Abstract) OR validate(Title/Abstract))

di, diagnosis subheading; exp, explode; fs, floating subheading; mp, multiple posting (term appears in title, abstract or MeSH); pt, publication type; sh, MeSH; :, truncation; tu, therapeutic use subheading; tw, textword; xs, exploded subheading.

table 5. In all cases, the sensitivities of the search filters did not differ statistically when comparing results derived in the 2000 clinical hedges database with those derived in a more current

Table 4 Number of citations included in the 2000 clinical hedges database and the McMaster PLUS databases (more current publishing years) by purpose category

Article category	2000 Clinical hedges database	McMaster PLUS database*
Therapy		
All articles	49 028	41 562
Methodologically rigorous	1587	1341
Diagnosis		
All articles	49 028	265 467
Methodologically rigorous	147	477
Prognosis		
All articles	49 028	265 467
Methodologically rigorous	190	868
Etiology		
All articles	49 028	89 214
Methodologically rigorous	281	550
Clinical prediction guides		
All articles	49 028	265 467
Methodologically rigorous	91	733

*Therapy—2010 database; Diagnosis, prognosis and clinical prediction guides—2003–10 database; Etiology—2007/8 database.

database (all p values >0.05). In addition, in all cases, the specificities of the search filters shown in table 5 remained well above 50% when testing them in the more current database.

The results for filters developed to maximize specificity for therapy, diagnosis, prognosis, etiology and clinical prediction guides are shown in table 6. In all cases, the specificities of the search filters differed statistically when comparing results derived in the 2000 clinical hedges database with those derived in a more current database (all p values <0.05). Although the results for specificity differed statistically (due to the large sample size), the differences are small and, thus, inconsequential. For example, for studies of therapy, prognosis and etiology, the specificities in the more current database were slightly lower than in the 2000 clinical hedges database, by 1.0%, 1.9% and 4.5%, respectively. However, in each of these instances, the sensitivities were slightly higher than in 2000. For studies of diagnosis the specificity in the current database was slightly higher, by 0.4%. For studies of clinical prediction guides the specificity and sensitivity in the current database were slightly lower, by 0.1% and 2.5%, respectively.

DISCUSSION

The results of this study show that the clinical queries that are available in the widely used PubMed and Ovid MEDLINE interfaces are robust when searching in more current publishing years. Estimates for therapy in the current database are similarly precise as those for 2000. Estimates for diagnosis, prognosis, etiology and clinical prediction guides are more precise for the current database because of its increased sample size.

Searchers can be assured that the clinical queries for therapy, diagnosis, prognosis, etiology, and clinical prediction guides perform equally well in detecting relevant, methodologically sound studies in the current publishing years as they did when they were derived over a decade ago in the publishing year 2000.

The reason for this stability is probably because little or no change has occurred during this time period in the basic methodological standards for conducting applied clinical research combined with no substantive changes in the indexing of the methodological features of these studies. In addition, no new

Table 5 Results for the testing of the sensitive search filters in the 2000 clinical hedges database and the more current database (therapy in a 2010 database; diagnosis, prognosis and clinical prediction guides in a 2003–10 database; etiology in a 2007–8 database)

Sensitive search filter	Sensitivity (%)	Specificity (%)
	2000 (95% CI)	2000 (95% CI)
	Current (95% CI)	Current (95% CI)
Therapy		
clinical trial.mp.	99.2 (98.7 to 99.6)	70.1 (69.7 to 70.6)
OR clinical trial.pt.	99.8 (99.5 to 100.0)	64.6 (64.2 to 65.1)
OR random:.mp.		
OR tu.xs.		
Diagnosis		
sensitiv:.mp.	98.6 (96.8 to 100.0)	74.9 (73.9 to 74.7)
OR diagnos:.mp.	98.3 (97.2 to 99.5)	75.9 (75.7 to 76.1)
OR di.fs.		
Prognosis		
incidence.sh.	86.8 (82.0 to 91.6)	79.7 (79.4 to 80.1)
OR exp mortality	83.9 (81.4 to 86.3)	78.7 (78.5 to 78.8)
OR follow-up studies.sh.		
OR prognos:.tw.		
OR predict:.tw.		
OR course:.tw.		
Etiology		
risk:.mp.	93.2 (90.3 to 96.2)	63.2 (62.7 to 63.6)
OR exp cohort studies	95.6 (93.9 to 97.3)	66.1 (65.8 to 66.4)
OR between group:.tw.		
Clinical prediction guides		
predict:.mp.	94.5 (89.8 to 99.2)	79.4 (79.1 to 79.8)
OR scor:.tw.	95.6 (94.2 to 97.1)	78.0 (77.9 to 78.2)
OR observ:.mp.		

di, diagnosis subheading; exp, explode; fs, floating subheading; mp, multiple posting (term appears in title, abstract or MeSH); pt, publication type; sh, MeSH; :, truncation; tu, therapeutic use subheading; tw, textword; xs, exploded subheading.

methodological terms relating to the design characteristics of high quality therapy, diagnosis, prognosis, etiology, and clinical prediction guide primary studies have been added to MeSH.

Clinical queries for the purpose categories indicated in this study as well as for other purpose categories (eg, qualitative

Table 6 Results for the testing of the specific search filters in the 2000 clinical hedges database and the more current database (therapy in a 2010 database; diagnosis, prognosis and clinical prediction guides in a 2003–10 database; etiology in a 2007–8 database)

Specific search filter	Sensitivity (%)	Specificity (%)
	2000 (95% CI)	2000 (95% CI)
	Current (95% CI)	Current (95% CI)
Therapy		
randomized controlled trial.pt.	94.0 (92.8 to 95.1)	97.5 (97.3 to 97.6)
OR randomized controlled trial.mp.	98.5 (97.9 to 99.2)	96.5 (96.3 to 96.7)
Diagnosis		
specificity.tw.	64.6 (56.9 to 72.4)	98.4 (98.2 to 98.5)
	73.6 (69.6 to 77.4)	98.8 (98.7 to 98.9)
Prognosis		
prognos:.tw.	50.0 (43.4 to 57.6)	94.1 (93.9 to 94.4)
OR first episode.tw.	53.6 (50.3 to 56.9)	92.2 (92.1 to 92.3)
OR cohort.tw.		
Etiology		
relative risk:.tw.	45.9* (40.1 to 51.7)	96.3 (96.2 to 96.5)
OR risks.tw.	57.8 (53.7 to 61.9)	91.8 (91.6 to 92.0)
OR cohort stud:.mp.		
Clinical prediction guides		
validation.tw.	53.4 (43.6 to 64.1)	99.3 (99.2 to 99.4)
OR validate.tw.	50.9 (47.2 to 54.5)	99.2 (99.1 to 99.3)

*We previously reported a sensitivity of 51% for the highly specific etiology search filter. The correct value is shown in this table, 45.9%. mp, multiple posting (term appears in title, abstract or MeSH); pt, publication type; :, truncation; tw, textword.

Table 7 Key aspects of the study

Key issues relating to search filter evaluation	Study reported in this article
Intended purpose/audience of the search filters	Clinicians—use of the specific filters when looking for a few good articles to answer patient care questions. Researchers—use of the sensitive filters when conducting an exhaustive search of the literature, for example, when conducting a systematic review.
Database interface used	Search filters were derived and tested in the Ovid MEDLINE database. Search filters were translated for use in PubMed. Translations were derived in collaboration with staff from the National Library of Medicine.
Methods of search term derivation	Clinical hedges database—to construct a comprehensive set of possible search terms, we listed MeSH terms and textwords related to study criteria and then sought input from clinicians and librarians through interviews, requests by email and at meetings and conferences, review of published and unpublished searching filters from other groups, and requests to MEDLINE experts. Individuals were asked what terms or phrases they used when searching for each study type. Terms could be subject headings, publication types, and subheadings, or could be single words or phrases as textwords, denoting their presence in titles and abstracts of articles. Various truncations were also applied to the textwords, phrases, and MeSH terms. We compiled a list of 5395 terms of which 4862 were unique. All terms were tested for all study types using the Ovid technologies searching system. The search filters derived in the clinical hedges database were tested and validated in more current gold standard databases (McMaster PLUS database).
Gold standard database for search filter derivation	Clinical hedges database—hand search of 161 journals that were indexed in MEDLINE.
Gold standard database for search filter validation	McMaster PLUS database—contains articles that meet criteria for scientific merit and clinical relevance in journals indexed by MEDLINE (up to 184 were hand searched during the time period in question). This database was augmented with all articles, from the same journals, that do not meet explicit criteria for scientific merit and clinical relevance.
Potential limitation of the search filters	The search filters were derived and validated using a set of journals that have been shown to publish high quality, clinically relevant content. The filters retrieve a very high proportion of all high quality studies and reduce the number of lower quality studies that must be reviewed by the user, but do not eliminate the need for critical appraisal.
Currency of the search filters	Validated in gold standard databases as recently as 2010.

studies) are also available for use in Ovid EMBASE, Ovid PsycInfo, and EBSCO CINAHL. We did not check the performance of these filters as they were calibrated in the 2000 database and have similar performance to the PubMed/MEDLINE clinical queries. On the basis of the current study, they are likely to have maintained their performance. Searchers interested in clinical topics that match the purpose categories in table 1 can use clinical queries filters to help limit the retrieval of articles to those that are scientifically sound and relevant to the methodological purpose of their search (eg, searching for studies about therapy in a particular content area, for example, asthma in children). As shown in table 4 the number of methodologically rigorous, clinically relevant studies in the two gold standard databases are extremely low, an observation that can be generalized to the MEDLINE database; underscoring the need to use the clinical queries. Sensitive filters will retrieve almost all sound studies, with fewer ‘false positive’ articles

than for traditional searching. Filters that are designed to maximize specificity will miss some 'true positive' articles but will retrieve far fewer false positive articles.

Information retrieval researchers can also use the databases used in this study to test their own methodologies and tools for search filter development and validation. As is the case for the clinical hedges database, the McMaster PLUS gold standard databases constructed as part of this study are available for use by other researchers. We have previously collaborated with researchers who have used the clinical hedges database for machine learning purposes. For example, we collaborated with Kilicoglu and colleagues²² at the NLM who approached the problem of recognizing studies containing useable clinical advice from retrieved topically relevant articles as a binary classification problem. The clinical hedges database formed the basis of their approach. They identified scientifically rigorous studies using supervised machine learning techniques (naive Bayes, support vector machine, and boosting) trained on high-level semantic features. They combined these methods using an ensemble learning method (stacking). In their study they used a training set of 10 000 manually annotated MEDLINE citations (from the clinical hedges database), and a test set of an additional 2000 citations (from the clinical hedges database). Their results demonstrated that a high quality gold standard and advanced classification methods can help clinicians acquire the best evidence from the medical literature. This is just one of many potential researcher usages of the clinical hedges database and the McMaster PLUS databases.

In conclusion, the clinical queries search filters have been revalidated and can remain a useful public resource for searching the world's medical literature for the research that is most relevant to clinical care. Key aspects of this study are summarized in table 7.

Acknowledgments The authors would like to acknowledge the computer programmers at Health Information Research Unit, Chris Cotoi and Nicholas Hobson.

Contributors All authors contributed to the conception, design, analysis and interpretation of the data. In addition, NLW acquired all the data. All authors were involved in drafting and reviewing the article for important intellectual content and all approved the version of the article to be published.

Funding This study was funded by the Canadian Institutes of Health Research (CIHR). CIHR had no involvement in the study design; in the collection, analysis or interpretation of the data; in the writing of the report; or in the decision to submit the paper for publication.

Competing interests None.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement Information retrieval researchers can use the databases used in this study to test their own methodologies and tools for search filter development and validation.

REFERENCES

1. **US National Library of Medicine.** National Institutes of Health. Fact Sheet MEDLINE. <http://www.nlm.nih.gov/pubs/factsheets/medline.html> (accessed 5 Apr 2012).
2. **Cullen RJ.** In search of evidence: family practitioners' use of the Internet for clinical information. *J Med Libr Assoc* 2002;**90**:370–9.
3. **Westbrook JI, Gosling AS, Coiera E.** Do clinicians use online evidence to support patient care? A study of 55,000 clinicians. *J Am Med Inform Assoc* 2004;**11**:113–20.
4. **Schwartz KL, Roe T, Northup J, et al.** Family medicine patients' use of the Internet for health information: a MetroNet study. *J Am Board Fam Med* 2006;**19**:39–45.
5. **Peterson G, Aslani P, Williams KA.** How do consumers search for and appraise information on medicines on the Internet? A qualitative study using focus groups. *J Med Intern Res* 2003;**5**:e33.
6. **US National Library of Medicine.** National Institutes of Health. Key MEDLINE Indicators. http://www.nlm.nih.gov/bsd/bsd_key.html (accessed 5 Apr 2012).
7. **Demner-Fushman D, Hauser SE, Humphrey SM, et al.** MEDLINE as a source of just-in-time answers to clinical questions. *AMIA Annu Symp Proc* 2006:190–4.
8. **Shariff SZ, Sontrop JM, Haynes RB, et al.** Impact of PubMed search filters on the retrieval of evidence by physicians. *Can Med Assoc J* 2012;**184**:E184–90.
9. **Schulz F, Altman DG, Moher D, for the CONSORT Group.** CONSORT 201 Statement: updated guidelines for reporting parallel group randomized trials. *BMJ* 2010;**340**:c332.
10. **Moher D, Liberati A, Tetzlaff J, et al. for the PRISMA Group.** Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *BMJ* 2009;**339**:b2535.
11. **Bossuyt PM, Reitsma JB, Bruns DE, et al.** Standards for reporting of diagnostic accuracy. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. *BMJ* 2003;**326**:41–4.
12. **Health Information Research Unit.** McMaster University. McMaster Health Knowledge Refinery (HKR). http://hiru.mcmaster.ca/hiru/HIRU_McMaster_HKR.aspx (Accessed 10 Apr 2012).
13. **Health Information Research Unit.** McMaster University. McMaster PLUS. http://hiru.mcmaster.ca/hiru/HIRU_McMaster_PLUS_projects.aspx (Accessed 10 Apr 2012).
14. **Health Information Research Unit.** McMaster University. Journal reviewed list. <http://hiru.mcmaster.ca/hiru/journalslist.asp> (Accessed 10 Apr 2012).
15. **Health Information Research Unit.** McMaster University. Inclusion criteria. <http://hiru.mcmaster.ca/hiru/InclusionCriteria.html> (Accessed 10 Apr 2012).
16. **Haynes RB, McKibbon KA, Wilczynski NL, et al. Hedges Team.** Optimal search strategies for retrieving scientifically strong studies of treatment from Medline: analytical survey. *BMJ* 2005;**330**:1179. Epub 13 May 2005.
17. **Haynes RB, Wilczynski NL.** Optimal search strategies for retrieving scientifically strong studies of diagnosis from Medline: analytical survey. *BMJ* 2004;**328**:1040. Epub 8 Apr 2004.
18. **Wilczynski NL, Haynes RB; Hedges Team.** Developing optimal search strategies for detecting clinically sound prognostic studies in MEDLINE: an analytical survey. *BMC Med* 2004;**2**:23.
19. **Wilczynski NL, Haynes RB; Hedges Team.** Developing optimal search strategies for detecting clinically sound causation studies in MEDLINE. *AMIA Annu Symp Proc* 2003:719–23.
20. **Wong SS, Wilczynski NL, Haynes RB, et al.; Hedges Team.** Developing optimal search strategies for detecting sound clinical prediction studies in MEDLINE. *AMIA Annu Symp Proc* 2003:728–32.
21. **Yao X, Wilczynski NL, Walter SD, et al.** Sample size determination for bibliographic retrieval studies. *BMC Med Inform Decis Mak* 2008;**8**:43.
22. **Kilicoglu H, Demner-Fushman D, Rindflesch TC, et al.** Towards automatic recognition of scientifically rigorous clinical research evidence. *J Am Med Inform Assoc* 2009;**16**:25–31.