Cite this article as: BMJ, doi:10.1136/bmj.38068.557998.EE (published 8 April 2004)

Papers

Optimal search strategies for retrieving scientifically strong studies of diagnosis from Medline: analytical survey

R Brian Haynes, Nancy L Wilczynski for the Hedges Team

Abstract

Objective To develop optimal search strategies in Medline for retrieving sound clinical studies on the diagnosis of health disorders.

Design Analytical survey.

Setting Medline, 2000.

Participants 170 journals for 2000 of which 161 were indexed in Medline.

Main outcome measures The sensitivity, specificity, precision ("positive predictive value"), and accuracy of 4862 unique terms in 17 287 combinations were determined by comparison with a hand search of all articles (the "gold standard") in 161 journals published during 2000 (49 028 articles).

Results Only 147 (18.9%) of 778 articles about diagnostic tests met basic criteria for scientific merit. Combinations of search terms reached peak sensitivities of 98.6% at a specificity of 74.3%. Compared with best single terms, best multiple terms increased sensitivity for sound studies by 6.8% (absolute increase), while also increasing specificity (absolute increase 6.0%) when sensitivity was maximised. When terms were combined to maximise specificity, the single term, specificity.tw. (98.4%), outperformed combinations of terms. The strategies newly reported in this paper outperformed other validated search strategies except for one strategy that had slightly higher sensitivity (99.3% *v* 98.6%) but lower specificity (54.7% *v* 74.3%).

Conclusion New empirical search strategies in Medline can optimise retrieval of articles reporting high quality clinical studies of diagnosis.

Introduction

Accurate diagnosis is the cornerstone of decision making for clinical intervention and is increasingly important as the number of validated treatments for specific conditions increases. Clinical research, usually widely accessible first in the biomedical journal literature, provides quantitative information about the sensitivity, specificity, and predictive value of many clinical and diagnostic tests, but this information is buried in a much larger biomedical literature. A recent survey showed that clinicians are highly interested in using evidence based information and frequently use Medline.[1] Information pertaining to diagnosis is second most commonly sought by clinicians after treatment.[2 3]

Finding the current best evidence in Medline for a diagnostic process is daunting, given that Medline has over 11 million articles from over 4500 journals, covering all aspects of biomedical and health research.[4] A recent qualitative study found that two of the six obstacles to answering clinical questions with evidence were the time required to find information and the difficulty in selecting an optimal search strategy.[5] Even clinicians who in principle support the use of evidence for patient care often do not have time to find and apply it in practice.[6] When they do try, searches are not performed effectively.[7]

Search filters ("hedges") can improve the retrieval of clinically relevant and scientifically sound studies from Medline and similar databases.[8–12] For instance, when we searched Medline for studies on the diagnosis of arthritis from 1996 to the present using the term "arthritis", 7083 articles alone were retrieved; using "arthritis and diagnosis" yielded 3451 articles. Although this filtered out over half the articles, there were still many articles to sort through, with no guarantee that the most rigorous studies would be retrieved. More sophisticated search filters can be created by combining disease content terms with medical subject headings, explosions, publication types, subheadings, and textwords (see box). These detect design features indicating methodological rigour for applied healthcare research using such terms as "gold standard" as a filter, seeking studies in which a test of uncertain value is compared with one of known high accuracy.

In the early 1990s our group at McMaster University developed search filters on a small subset of 10 journals and for four types of article (therapy, diagnosis, prognosis, and causation (aetiology)).[13 14] These strategies have been adapted for use in the Clinical Queries interface of Medline (www.ncbi.nlm.nih.gov/entrez/query/static/clinical.html). This research is being updated and expanded with data from 161 journals indexed in Medline from 2000. The robustness of empirical search strategies developed in 1991 for detecting clinical content in Medline in 2000 has already been reported.[15] We report on the information retrieval properties of single terms and combinations of terms in Medline for identifying methodologically sound studies on the diagnosis of health disorders.

Methods

We developed search strategies by using methodological search terms and phrases in a subset of Medline records matched with a handsearch of the contents of 161 journal titles for 2000. The search strategies were treated as diagnostic tests for sound studies, and the manual review of the literature was treated as the gold standard. It is potentially confusing to use the terminology of diagnostic testing for assessing strategies for retrieving articles about diagnostic tests, especially when some of the search terms are the same. Nevertheless, the principles for retrieval are the

P+ *Full titles of journals indexed in Medline and a specific diagnostic search strategy are on bmj.com*

Table 1 Formula for calculating sensitivity, specificity, precision, and accuracy of Medline searches for detecting sound studies of diagnosis by manual review

Search terms	Meets criteria	Does not meet criteria
Terms detected	a	b
Terms not detected	c	d

Sensitivity=a/(a+c); precision=a/(a+b); specificity=d/(b+d); accuracy=(a+d)/(a+b+c+d). All articles classified during manual review of literature=(a+b+c+d).

same as those for diagnosis. Thus we determined the sensitivity, specificity, accuracy, and precision (a library science term equivalent to the diagnostic test term "positive predictive value") of single term and multiple term Medline search strategies (table 1 and box). Sensitivity and specificity are not affected by the proportion of high quality articles in the database; precision depends on this proportion, and so does accuracy, but to a lesser extent.

After extensive attempts only 2% (n = 968) of the handsearch items did not match citations in Medline. Unmatched citations that were detected by a search strategy were included in cell b of the analysis table (table 1), leading to slight underestimates of the precision, specificity, and accuracy of the search strategy. Similarly, unmatched citations that were not detected by a search strategy were included in cell d of the table, leading to slight overestimates of specificity and accuracy.

Manual review

Six research assistants reviewed all issues of 170 journals for 2000 of which 161 were indexed in Medline. The journal titles were regularly reviewed for content for four evidence based journals prepared by our group, *Evidence-Based Medicine*, *Evidence-Based Nursing*, *Evidence-Based Mental Health*, and *ACP Journal Club*, according to an explicit process that assesses the scientific merit and clinical relevance of original and review articles for health care (www.acpjc.org/shared/purpose_and_procedure.htm). The journal list has been chosen over several years in an iterative process based on handsearch review of over 400 journals recommended by clinicians and librarians, science citation index impact factors, recommendations by editors and publishers, and ongoing assessment of their yield of studies and reviews of scientific merit and clinical relevance. These journals (examples bracketed) include content for the disciplines of internal medicine (*Annals of Internal Medicine*), general medical practice (*BMJ, JAMA*, and *Lancet*), mental health (*Archives of General Psychiatry, British Journal of Psychiatry*), and general nursing practice (*Nursing Research*) (also see bmj.com).

Terms and definitions for search strategies

- Sensitivity—proportion of high quality articles retrieved
- Specificity—proportion of low quality diagnosis studies or non-diagnosis studies not retrieved
- Precision—proportion of retrieved articles of high quality
- Accuracy—proportion of all articles correctly categorised
- "ANDed"—combined with
- di—diagnosis subheading
- du—diagnostic use subheading
- exp—explosion
- fs—floating subheading
- MeSH—medical subject heading
- mp—multiple posting (term in title, abstract, or MeSH heading)
- pt—publication type
- sh—MeSH subject heading
- tw—textword
- xs—exploded subheading
- :—truncation

Methodological criteria for evaluating studies of diagnosis were: inclusion of a range of participants; use of an objective diagnostic ("gold") standard or current clinical standard for diagnosis; participants receiving the new test and some form of the diagnostic standard; interpretation of diagnostic standard without knowledge of test result, and vice versa; and analysis consistent with study design. These criteria were developed for critical appraisal of the healthcare literature, and the second to fourth criteria have been empirically validated.[16 17] The research assistants were rigorously calibrated and periodically checked for application of criteria to determine if each article was methodologically sound for any of six categories of purpose (diagnosis and screening, treatment and prevention, prognosis, aetiology and harm, clinical prediction guides, and economics).[18] Inter-rater agreement for identifying the purpose of articles was 81% beyond chance (κ 0.81, 95% confidence interval 0.79 to 0.84). Inter-rater agreement for which articles met all scientific criteria was 89% beyond chance (κ 0.89, 0.78 to 0.99).[18] Articles that seemed to pass the criteria were reviewed by at least the lead author (RBH).

Collecting search terms

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 strategies from other groups, and requests to Medline experts. Individuals were asked what terms or phrases they used when searching for each category. Terms could be subject headings, publication types, check tags, 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 in all purpose categories using the Ovid Technologies searching system. Optimised strategies for aetiology and studies of clinical prediction guides have been published elsewhere.[19 20]

Data collection

Data collection forms were used to record handsearched data for each article found in each issue of the 161 journal titles. These data were scanned using Teleform software (Cardiff Software; Vista, CA). After verification of the data online, the handsearch data were written to an Access database (Microsoft). Each journal title was searched in Medline for 2000, and the full Medline records were captured for all articles in the journals. Medline data were then linked with the handsearch data.

Testing strategies

We calculated the sensitivity, specificity, precision, and accuracy for each term for each category of article. For some categories of articles, such as therapy, we were able to split the database into 60% and 40% components to provide a development and validation database. For diagnosis, however, this was not possible as there were an insufficient number of diagnosis articles that were considered methodologically rigorous. Individual search terms with a sensitivity of more than 25% and a specificity of more than 75% for the diagnosis category were incorporated into the development of search strategies that included a combi-

Table 2 Best single terms for high sensitivity searches, high specificity searches, and searches that optimise the balance between sensitivity and specificity for retrieving studies of diagnosis. Values are percentages (95% confidence intervals)

Search strategy in Ovid format	Sensitivity (n=147)	Specificity (n=48 881)	Precision*	Accuracy (n=49 028)
High sensitivity†: di.xs.	91.8 (87.4 to 96.3)	68.3 (67.9 to 68.7)	0.9 (0.7 to 1.0)	68.4 (68.0 to 68.8)
High specificity‡: specificity.tw.	64.6 (56.9 to 72.4)	98.4 (98.2 to 98.5)	10.6 (8.6 to 12.6)	98.3 (98.1 to 98.4)
Optimising sensitivity and specificity§: exp "diagnostic techniques and procedures"	66.7 (59.1 to 74.3)	74.6 (74.2 to 75.0)	0.8 (0.6 to 0.9)	74.5 (74.2 to 74.9)

See box for description of terms.
*Denominator varies by row; see table 1 for calculation.
†Keeping specificity ≥50%.
‡Keeping sensitivity ≥50%.
§Keeping (abs(sensitivity–specificity)) to a minimum.

Table 3 Top three search strategies yielding highest sensitivity (keeping specificity ≥50%) with combinations of terms. Values are percentages (95% confidence intervals)

Search strategy in Ovid format	Sensitivity (n=147)	Specificity (n=48 881)	Precision*	Accuracy (n=49 028)
sensitiv:.mp. OR diagnos:.mp. OR di.fs.	98.6 (96.8 to 100.0)	74.3 (73.9 to 74.7)	1.1 (1.0 to 1.3)	74.3 (74.0 to 74.7)
sensitiv:.mp. OR diagnos:.mp. OR accuracy.tw.	98.0 (95.7 to 100.0)	82.7 (82.4 to 83.1)	1.7 (1.4 to 2.0)	82.8 (82.5 to 83.1)
sensitiv:.mp. OR diagnos:.mp. OR test:.tw.	98.0 (95.7 to 100.0)	75.1 (74.8 to 75.5)	1.2 (1.0 to 1.4)	75.2 (74.8 to 75.6)

See box for description of terms.
*Denominator varies by row; see table 1 for calculation.

nation of two or more terms. All combinations of terms used the Boolean OR—for example, "sensitivity OR specificity".

For the development of multiple term search strategies to optimise either sensitivity or specificity, we tested the combination of individual terms with all two term search strategies with sensitivity at least 75% and specificity at least 50%. For optimising accuracy, two term search strategies with accuracy of more than 75% were considered for multiple term development. Overall, we tested 17 287 multiple term search strategies. Search strategies were also developed that optimised combined sensitivity and specificity (equivalent to the optimal point on a receiver operating characteristic curve, minimising the total number of errors).

Results

Overall, 49 028 articles were included in the analysis. Of these, 778 (1.6% of original studies and review articles, case reports, or general interest papers) were classified as original studies evaluating a diagnosis question, of which 147 (18.9%) met the methodological criteria.

Table 2 shows the operating characteristics for the single terms with the highest sensitivity and specificity. The best accuracy when keeping sensitivity to 50% or more was seen with the term "specificity.tw." (.tw. is Ovid search system's syntax for searching all words in the title and abstract of an article).

Tables 3 and 4 show the strategies yielding the highest sensitivity and specificity based on testing of all strategies for combinations up to three terms. Some one term and two term strategies outperformed multiple term strategies (table 4). Because of the low prevalence of diagnosis articles, the accuracy of search terms is driven by their specificity, and thus the three search strategies yielding the highest accuracy are the same as those yielding the highest specificity (table 4). Table 5 shows the three search strategies best optimising the trade off between sensitivity and specificity.

Logistic regression modelling did not lead to the development of search strategies that outperformed those already developed using the Boolean approach.

We used our data to test 10 published strategies and one previously unpublished strategy for retrieving diagnostic test studies from Medline.[9–11] Two strategies were modified slightly to eliminate the content words in the search strategies. When we used our handsearch data, the published and unpublished strategies containing only methodological terms had a sensitivity range of 85.0% to 99.3%. One strategy had slightly higher sensitivity (99.3%) than our most sensitive strategy (98.6%), but it came with a large trade off for specificity (54.7%, compared with

Table 4 Top three search strategies yielding highest specificity (and highest accuracy) (keeping sensitivity ≥50%) with combinations of terms. Values are percentages (95% confidence intervals)

Search strategy in Ovid format	Sensitivity (n=147)	Specificity (n=48 881)	Precision*	Accuracy (n=49 028)
specificity.tw.	64.6 (56.9 to 72.4)	98.4 (98.2 to 98.5)	10.6 (8.6 to 12.6)	98.3 (98.1 to 98.4)
specificity.tw. OR predictive value:.tw.	72.8 (65.6 to 80.0)	97.9 (97.8 to 98.1)	9.6 (7.9 to 11.3)	97.9 (97.7 to 98.0)
accurac:.tw. OR predictive value:.tw.	52.4 (44.3 to 60.5)	97.9 (97.8 to 98.1)	7.1 (5.6 to 8.6)	97.8 (97.7 to 97.9)

See box for description of terms.
*Denominator varies by row; see table 1 for calculation.

Table 5 Top three search strategies for optimising sensitivity and specificity (based on minimising absolute difference between sensitivity and specificity). Values are percentages (95% confidence intervals)

Search strategy using Ovid format	Sensitivity (n=147)	Specificity (n=48 881)	Precision*	Accuracy (n=49 028)
sensitiv:.mp. OR predictive value:.mp. OR accurac:.tw.	92.5 (88.3 to 96.8)	92.1 (91.8 to 92.3)	3.4 (2.8 to 3.9)	92.1 (91.8 to 92.3)
sensitiv:.mp. OR predictive value:.mp. OR accuracy.tw.	92.5 (88.3 to 96.8)	92.1 (91.8 to 92.3)	3.4 (2.8 to 3.9)	92.1 (91.8 to 92.3)
sensitiv:.mp. OR diagnostic.mp. OR predictive value:.tw.	92.5 (88.3 to 96.8)	91.8 (91.6 to 92.1)	3.3 (2.8 to 3.8)	91.8 (91.6 to 92.1)

See box for description of search terms.
*Denominator varies by row; see table 1 for calculation.

Table 6 Comparison of performance of strategies from 1991 and 2000, compiled using 2000 dataset. Values are percentages

Approach and year	Strategy in Ovid format	Sensitivity (n=147)	Specificity (n=48 881)	Precision*	Accuracy (n=49 028)
Maximise sensitivity:					
1991	exp sensitivity and specificity OR sensitivity.tw. OR di.xs. OR du.fs. OR specificity.tw.	96.6	65.0	0.8	65.7
2000	sensitiv:.mp. OR diagnos:.mp. OR di.fs.	98.6	74.3	1.1	74.3
Maximise specificity:					
1991	exp sensitivity and specificity OR predictive value:.tw.	79.6	94.9	4.5	94.8
2000	specificity.tw.	64.6	98.4	10.6	98.3

See box for description of terms.

*Denominator varies by row; see table 1 for calculation.

our strategy's specificity of 74.3%; see table 3). The specificities for these strategies in our database ranged from 54.7% to 94.5%, all lower than our best specificity of 98.4% (see table 4).

Discussion

Our study documents search terms with best sensitivity, specificity, accuracy, and balance of sensitivity and specificity for retrieving high quality studies of diagnostic tests from Medline. This research updates our previous one published in 1994, calibrated using 10 internal and general medicine journals.[19] When the 1991 strategies for diagnosis articles were tested in the 2000 database, the performance of the 2000 strategies was consistently better (table 6). We did not have enough data to do an independent validation of our diagnostic test strategies and thus risked overestimating their performance. We did independent validations for studies of therapy, however, with the greatest statistically significant difference being 1.1% for one set of specificities (data not shown). Furthermore, by double checking only articles that initially seemed to pass criteria, we may have underestimated performance: a few articles that met our criteria may have been missed in the handsearch.

Searchers who want retrieval with little non-relevant material can choose strategies with high specificity. For those interested in comprehensive retrievals or in searching for clinical topics with few citations, strategies with higher sensitivity may be more appropriate. The strategies that optimised the balance of sensitivity and specificity provided the best separation of eligible studies from others but did so without regard for whether sensitivity or specificity was affected. Regardless of the strategy used, we foresee that the most effective way to harness these strategies is to have them embedded within searching systems, either as clinical queries in PubMed or as stored searches that can be invoked at the user's request. The US National Library of Medicine has updated their Clinical Queries site for searching Medline for studies of diagnostic tests and other clinical topics, and they are available free (web.ncbi.nlm.nih.gov/entrez/query/static/clinical.shtml). Further, the new strategies have been incorporated into Ovid's main search engine for Medline (www.ovid.com), with the high specificity strategies being incorporated into Skolar (www.skolar.com).

Our search strategies were designed to retrieve diagnostic test studies that meet criteria for validity, just 18.9% of all diagnosis studies in our database. We did not test the performance of these strategies for all diagnosis studies, but in a similar project for studies of health services research, we found that the highest sensitivity strategies for the better designed studies had 5-10% lower sensitivity for all articles on the same topic, with no important differences in specificity (unpublished data).

Other investigators have attempted to find strategies that outperform those we previously published, with some success.[9-12 14] Our new strategies have set the bar higher, but there is still considerable room for improvement, particularly for the precision of searches.

What is already known on this topic

Information on the accuracy of diagnostic tests abounds in the medical literature but is often unknown to, or forgotten by, clinicians

The medical literature is accessible through large internet databases such as Medline, but few clinicians know how to search them well

What this study adds

Special Medline search strategies were developed and tested that retrieved up to 99% of scientifically strong studies of diagnostic tests

These strategies have been automated for use in PubMed Medline at a special screen, Clinical Queries, and Ovid Technology's Medline and Skolar services

The Hedges Team includes Angela Eady, Brian Haynes, Susan Marks, Ann McKibbon, Doug Morgan, Cindy Walker-Dilks, Stephen Walter, Stephen Werre, Nancy Wilczynski, and Sharon Wong, all at McMaster University Faculty of Health Sciences.

Contributors: RBH planned the study, designed the protocol, and interpreted the data; he will act as guarantor. NLW supervised the research staff, and collected, analysed, and interpreted the data. The Hedges Team conducted the study: AE, SM, AM, CW-D, S Werre, and S Wong collected the data. DM programmed the data set and analysed the data. S Walter and S Werre provided statistical advice, and S Werre did supplementary analyses. The manuscript was prepared by NLW and RBH.

Funding: This study was funded by the US National Institutes of Health (grant No 1 RO1 LM06866).

Competing interests: None declared.

Ethical approval: Not required.

1 Veness M, Rikard-Bell G, Ward J. Views of Australian and New Zealand radiation oncologists and registrars about evidence-based medicine and their access to internet based sources of evidence. *Australas Radiol* 2003;47:409-15.
2 Green ML, Ciampi MA, Ellis PJ. Residents' medical information needs in clinic: are they being met? *Am J Med* 2000;109:218-23.
3 Haynes RB, McKibbon KA, Walker CJ, Ryan N, Fitzgerald D, Ramsden MF. Online access to MEDLINE in clinical settings. A study of use and usefulness. *Ann Intern Med* 1990;112:78-84.
4 NLM Fact Sheet. www.nlm.nih.gov/pubs/factsheets/bsd.html (accessed 7 Dec 2003).
5 Ely JW, Osheroff JA, Ebell MH, Chambliss ML, Vinson DC, Stevermer JJ, et al. Obstacles to answering doctors' questions about patient care with evidence: qualitative study. *BMJ* 2002;324:710.
6 Tomlin Z, Humphrey C, Rogers S. General practitioners' perceptions of effective health care. *BMJ* 1999;318:1532-5.
7 Balas EA, Stockham MG, Mitchell JA, Sievert ME, Ewigman BG, Boren SA. In search of controlled evidence for health care quality improvement. *J Med Syst* 1997;21:21-32.
8 Haynes RB, McKibbon KA, Fitzgerald D, Guyatt GH, Walker CJ, Sackett DL. How to keep up with the medical literature. V. Access by personal computer to the medical literature. *Ann Intern Med* 1986;105:810-6.
9 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.

10 Deville 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.
11 Van der Weijden T, Ijzermans CJ, Dinant GJ, van Duijn NP, de Vet R, Buntinx F. Identifying relevant diagnostic studies in MEDLINE. The diagnostic value of the erythrocyte sedimentation rate (ESR) and dipstick as an example. *Fam Pract* 1997;14:204-8.
12 Vincent S, Greenley S, Beaven O. Clinical cvidence 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.
13 Wilczynski NL, Walker CJ, McKibbon KA, Haynes RB. Assessment of methodologic search filters in MEDLINE. *Proc Annu Symp Comput Appl Med Care* 1993;:601-5.
14 Haynes RB, Wilczynski N, McKibbon KA, Walker CJ, Sinclair JC. Developing optimal search strategies for detecting clinically sound studies in MEDLINE. *J Am Med Inform Assoc* 1994;1:447-58.
15 Wilczynski NL, Haynes RB, Hedges Team. Robustness of empirical search strategies for clinical content in MEDLINE. *Proc AMIA Symp* 2002;:904-8.
16 Jaeschke R, Guyatt GH, Sackett DL et al for the Evidence Based Medicine Working Group. Users' guides to the medical literature: III-How to use an article about a diagnostic test A. Are the results of the study valid? *JAMA* 1994;271:389-91.
17 Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, van der Meulen JH, et al. Empirical evidence of design-related bias in studies of diagnostic tests. *JAMA* 1999;282:1061-6.
18 Wilczynski NL, McKibbon KA, Haynes RB. Enhancing retrieval of best evidence for health care from bibliographic databases: calibration of the hand search of the literature. *Medinfo* 2001;10(Pt 1):390-3.
19 Wilczynski NL, Haynes RB for the Hedges Team. Developing optimal search strategies for detecting clinically sound causation studies in MEDLINE. *Proc AMIA Annu Symp* 2003:719-23.
20 Wong SSL, Wilczynski NL, Haynes RB, Ramkissoonsingh R for the Hedges Team. Developing optimal search strategies for detecting sound clinical prediction studies in MEDLINE. *Proc AMIA Annu Symp* 2003:728-32.

(Accepted 18 March 2004)

doi 10.1136/bmj.38068.557998.EE

Health Information Research Unit, Department of Clinical Epidemiology and Biostatistics, McMaster University Faculty of Health Sciences, 1200 Main Street West, Hamilton, ON, L8N 3Z5, Canada
R Brian Haynes *professor*
Nancy L Wilczynski *research associate*

Correspondence to: R B Haynes bhaynes@mcmaster.ca