Research Paper ■

Identifying Diagnostic Studies in MEDLINE: Reducing the Number Needed to Read

LUCAS M. BACHMANN, MD, RETO CORAY, MD, PIUS ESTERMANN, MD, GERBEN TER RIET, MD, PHD

Abstract **Objectives**. The search filters in PubMed have become a cornerstone in information retrieval in evidence-based practice. However, the filter for diagnostic studies is not fully satisfactory, because sensitive searches have low precision. The objective of this study was to construct and validate better search strategies to identify diagnostic articles recorded on MEDLINE with special emphasis on precision.

Design. A comparative, retrospective analysis was conducted. Four medical journals were hand-searched for diagnostic studies published in 1989 and 1994. Four other journals were hand-searched for 1999. The three sets of studies identified were used as gold standards. A new search strategy was constructed and tested using the 1989-subset of studies and validated in both the 1994 and 1999 subsets. We identified candidate text words for search strategies using a word frequency analysis of the abstracts. According to the frequency of identified terms, searches were run for each term independently. The sensitivity, precision, and number needed to read (1/precision) of every candidate term were calculated. Terms with the highest sensitivity × precision product were used as free text terms in combination with the MeSH term "SENSITIVITY AND SPECIFICITY" using the Boolean operator OR. In the 1994 and 1999 subsets, we performed head-to-head comparisons of the currently available PubMed filter with the one we developed.

Measurements. The sensitivity, precision and the number needed to read (1/precision) were measured for different search filters.

Results. The most frequently occurring three truncated terms (diagnos*; predict* and accura*) in combination with the MeSH term "SENSITIVITY AND SPECIFICITY" produced a sensitivity of 98.1 percent (95% confidence interval: 89.9–99.9%) and a number needed to read of 8.3 (95% confidence interval: 6.7–11.3%). In direct comparisons of the new filter with the currently available one in PubMed using the 1994 and 1999 subsets, the new filter achieved better precision (12.0% versus 8.2% in 1994 and 5.0% versus 4.3% in 1999. The 95% confidence intervals for the differences range from 0.05% to 7.5% (p = 0.041) and –1.0% to 2.3% (p = 0.45), respectively). The new filter achieved slightly better sensitivities than the currently available one in both subsets, namely 98.1 and 96.1% (p = 0.32) versus 95.1 and 88.8% (p = 0.125).

Conclusions. The quoted performance of the currently available filter for diagnostic studies in PubMed may be overstated. It appears that even single external validation may lead to over optimistic views of a filter's performance. Precision appears to be more unstable than sensitivity. In terms of sensitivity, our filter for diagnostic studies performed slightly better than the currently available one and it performed better with regards to precision in the 1994 subset. Additional research is required to determine whether these improvements are beneficial to searches in practice.

■ **J Am Med Inform Assoc.** 2002;9:653–658. DOI 10.1197/jamia.M1124.

Affiliations of the authors: University of Zürich, Zürich, Switzerland (LMB, RC, PE); University of Amsterdam (GtR).

Correspondence and reprints: Lucas M. Bachmann, MD, 1 Horten Centre, University of Zürich, Bolleystrasse 40, Postfach Nord, CH-8091 Zürich, Switzerland; e-mail: <lucas.bachmann@evimed.ch>.

Received for publication: 3/15/02; accepted for publication: 6/25/02.

Biomedical databases are important sources of evidence in medical practice. However, information retrieval in such databases can become very time-consuming because searches that are likely to identify all relevant information also find many irrelevant articles.

In recent years researchers have adopted various approaches in the development of search strategies to selectively retrieve different types of studies (therapy, prognosis, diagnosis and etiology) and different study designs.[1,2] Search strategies targeted at diagnostic studies have also been developed.[1,3–4] The most commonly used filter for diagnostic studies is almost certainly the one now publicly available in PubMed (Clinical Queries),[5] which based on the work of Haynes and coworkers.[1] Their search filter with emphasis on sensitivity achieved a cross-validated mean sensitivity of approximately 87% combined with a (non–cross-validated) mean precision of approximately 8%.

Compared with the filter for therapeutic studies, the diagnostic filter's precision in particular is much lower. The main reason for this difference may be the inconsistent terminology used in diagnostic studies making them difficult to index and retrieve in electronic databases.

In view of this high false-positive rate, we wondered if it would be possible to develop a more precise search strategy for selecting publications on diagnostic test evaluations without losing sensitivity. Our objective was to develop, test and validate a generic search strategy for the detection of diagnostic articles recorded on MEDLINE that can be applied in any diagnostic field in Medicine.

Methods

Two investigators (RC, LMB) independently hand-searched all issues published in 1989 and 1994 of the *European Journal of Paediatrics, Gastroenterology, American Journal of Obstetrics and Gynecology*, and *Thorax*. The journals used in this study are indexed cover to cover in MEDLINE. To obtain another validation set consisting not only of different years but also of different journals, all issues published in 1999 of the *New England Journal of Medicine, JAMA, BMJ*, and *The Lancet* were also similarly hand-searched. Diagnostic studies were defined as content pertained directly to the evaluation of a disease process, usually through comparing methods of arriving at a diagnosis."[1] A test was defined as any procedure used to change the estimate of the likelihood of disease presence. This includes components of history taking and physical examination and more technically advanced tests. Discrepancies between the two investigators were discussed and resolved by consensus. Only references of the diagnostic studies identified were stored in a Reference Manager file[6] and constituted our gold standard.

The gold standard references were identified in MEDLINE (Datastar Version) using the accession number, which is a unique identifier for a specific record. A strategy combining all accession numbers using the Boolean connector OR was saved. Thus, a search in MEDLINE would uniquely identify the gold standard references. To construct the search filter the number of references in MEDLINE was reduced to the subset of all references (1729) that were published in the four chosen journals in 1989. To validate the filter, MEDLINE references were reduced to the subset of all references (1797) that were published in the four journals in 1994. This approach was chosen to simulate a universe of searchable articles. To validate the filter under still stricter conditions it was tested in the 1999 subset, which consisted of different journals.

Selection of Free Text Terms

We applied the method of Boynton and coworkers[7]; that is, we selected potentially useful text words through a word frequency analysis. We performed the frequency analysis of the occurrence of each word in each record of the 1989-gold standard using the ListIndex function of the Idealist bibliographic software package.[8] Thus, we determined the frequency of all the words in the titles, abstracts, and subject indexes.

The list was transferred to an Excel file (MS Office 2000, Redmond, Washington 98052). To specifically select terms semantically associated with diagnosis, two investigators (PE and LMB) excluded numbers, single letters, author names and institutions, register numbers and journal names. Terms were also excluded if they were general medical language—for example, organ names or diseases, population of interest, or the word "study." We considered that these words would not be helpful in focusing a search on diagnostic studies. If the two investigators disagreed on excluding a term, it was included. All included expressions were sorted alphabetically. When terms differed only in the ending (e.g., diagnosis, diagnose, diagnostic, diagnostics) we decided to use the truncated term (e.g., "diagnos*"). With the 20 most prevalent (truncated) terms searches were run using each term independently (Table 1). Sensitivity (the number of retrieved gold standard articles as a proportion of *all* gold standard articles), precision (the number of gold standard articles as a proportion of all retrieved articles), and number needed to read (NNR = 1/precision) of each text word were then calculated. We coined the term NNR in analogy to the number needed to treat (NNT) to describe the number of irrelevant references that one has to screen to find one of relevance.

Next, the product of sensitivity and precision was computed for each of the text words as a single measure that strikes a balance between sensitivity and precision.

Construction of the Search Strategy

The terms with the highest sensitivity × precision product were combined in a stepwise fashion with the exploded MeSH term "SENSITIVITY AND SPECIFICITY" using OR. The sensitivities and precisions of these cumulative search strategies were then calculated to find the optimal search strategy (highest sensitivity in combination with highest precision). The term sensitiv* ranked third on the list of highest sensitivity × precision products. Since combination of this term with the MeSH term "SENSITIVITY AND SPECIFICITY" would have only contributed to an increase of sensitivity at the cost of precision, we also ran a search dropping this term but adding the next term (accura*) on the list.

Validating the Search Strategy

The search strategy with the highest sensitivity combined with the highest precision based on the set of 1989 gold standard articles was retested on the sets of 1994 and 1999 gold standard articles. We ran the search filter of Haynes et al.[1] in the 1994 and 1999 sets to compare its performance directly with the new filter.

Statistical Analysis

All calculations were performed with Stata software package (version 7.0, StataCorp. 1999. Stata Statistical Software: Relcase 7.0 College Station, Texas, USA). Ninety-five percent confidence intervals around single proportions were calculated using exact methods (Stata command: cii). Sensitivities were compared using McNemar's test using exact calculations for the p-values (Stata command: mcci). Precisions were compared using chi-squared tests for independent proportions using large-sample statistics (Stata command: prtcsti).

Table 1 ■

List of 20 (Truncated) Terms with Corresponding Sensitivities, Precisions and the Sensitivity × Precision Products when Searched as a Single Term

Term (truncated)	Sensitivity (%)	Precision (%)	Product	(Ranking)
predict*	48.2	36.4	1754.48	(1)
diagnos*	80.7	16.8	1355.76	(2)
sensitiv*	36.1	33.0	1191.3	(3)
accura*	24.1	46.5	1120.65	(4)
screen*	19.3	39.0	752.7	(5)
specific*	35.0	19.9	696.5	(6)
test*	49.4	13.7	676.78	(7)
detect*	32.5	18.1	588.25	(8)
positiv*	28.9	20.2	583.78	(9)
negativ*	20.5	20.0	410	(10)
evaluat*	27.7	11.6	321.32	(11)
analy*	38.6	7.7	297.22	(12)
risk*	22.9	12.8	293.12	(13)
assess*	24.1	11.6	279.56	(14)
scor*	13.3	17.2	228.76	(15)
assay*	13.3	16.7	222.11	(16)
differen*	28.9	7.5	216.75	(17)
measure*	25.3	7.6	192.28	(18)
examin*	20.5	8.4	172.2	(19)
determ*	22.9	7.4	169.46	(20)

Results

Eighty-three, 53, and 61 articles (gold standard) on diagnostics were identified out of 1729, 1797, and 7936 references (in the four journals) in 1989, 1994, and 1999, respectively. The 20 truncated terms with the highest frequency according to the ListIndex function (Idealist)[8] are listed in Table 1. The calculation of the sensitivity × precision products led to a new order of terms. The consecutive connection of these terms with the Boolean operator OR produced the final set of search strategies. Their performance is shown in Table 2.

The search strategy "SENSITIVITY AND SPECIFICITY"# OR predict* OR diagnos* OR sensitiv* resulted in a sensitivity of 92.8% (95% confidence interval: 84.9–97.3) and a NNR of 6.4 (95% CI: 5.2–8.0). In other words, approximately six abstracts have to be read to identify one on diagnostics. The search strategy "SENSITIVITY AND SPECIFICITY"# OR predict* OR diagnos* OR accura*, which ignored the truncated free text term "sensitiv," resulted in a sensitivity of 95.2% (95% confidence interval: 88.1–98.7) and a NNR of 5.9 (95% CI: 4.8–7.3). Based on its better performance we decided to validate the latter strategy in the 1994 and 1999 gold standard sets.

Table 2 ■

Development of Two Search Strategies with Stepwise Adding of Terms*

Search Strategy	Summary Performance Sensitivity (%)	Summary Performance Precision (%)	Number Needed to Read (NNR)
1989 test set (n = 1729)			
"SENSITIVITY AND SPECIFICITY"# OR predict* OR diagnos* OR sensitiv*	92.8	15.6	6.4
"SENSITIVITY AND SPECIFICITY"# OR predict* OR diagnos* OR accura*	95.2	16.9	5.9
1994 validation set (n = 1797)			
"SENSITIVITY AND SPECIFICITY"# OR predict* OR diagnos* OR sensitiv*	98.1	10.9	9.2
"SENSITIVITY AND SPECIFICITY"# OR predict* OR diagnos* OR accura*	98.1	12.0	8.3
1999 validation set (n = 7936)			
"SENSITIVITY AND SPECIFICITY"# OR predict* OR diagnos* OR sensitiv*	91.8	4.7	21.3
"SENSITIVITY AND SPECIFICITY"# OR predict* OR diagnos* OR accura*	95.1	5.0	20.0

*Terms were ranked according to their sensitivity × precision product. The Number Needed to Read figure shows how many abstracts have to be read to identify one diagnostic study and is equivalent to 1/precision.

Validation

In the 1994 subset, the new filter achieved a sensitivity of 98.1% (95% confidence interval: 89.9–99.9), a precision of 12% (95% confidence interval: 9.1–15.4) and a NNR of 8.3 (95% confidence interval: 6.7–11.3). The performance of the strategy "SENSITIVITY AND SPECIFICITY"# OR predict* OR diagnos* OR sensitiv* was slightly worse for precision (10.9%) and identical for sensitivity (98.1%). In the 1999 subset, which consisted of four other journals, the new filter retained its high sensitivity (95.1%), but precision was worse (5.0%). Table 3 provides the details for implementation of the search strategies in four commonly used MEDLINE interfaces.

Comparison with the Currently Available Filter

The currently available "optimal-sensitivity filter" has a quoted sensitivity of 92%. However, its true value may actually be 86%, since this was the figure found when this filter, which was derived in a 1991 data set, was run in the independent data set of 1986. We calculated its 95% confidence interval based on the data in Table 5 of Haynes et al.'s original paper. It ranges from 77.0 to 92.3%. Haynes filter's precision is quoted as 9%, but this figure has not been reproduced in an independent data set. We judged it safer to use the mean of the values found in the 1986 and 1991 data sets, which is 8%. Data to calculate confidence intervals are not in the original paper.

In a direct comparison of our filter with the currently available one in PubMed using the 1994 subset, the former achieved better precision (12.0% versus 8.2%; 95% CI of the difference: 0.05–7.5%; p = 0.041), although the sensitivities were almost identical, 98.1% and 96.2% for the new versus the current one, respectively. In a second direct comparison in which the new filter was tested under conditions that may theoretically have been more difficult (using other journals), it tended to outperform the currently available filter on sensitivity and precision, although the differences were not statistically significant. For the new filter, sensitivity was 95.1% against 88.5% for the current one; the difference was 6.6% (95% CI: V12.9–14.4%). For the new filter, precision was 5.0% against 4.3%; the difference was 0.7% (95%CI: V1.0–2.3%).

Discussion

In the table for clinical queries using research methodology filters in PubMed, one finds a summary of the characteristics of the PubMed filter for diagnostic studies. Using the sensitive filter we may expect to find between 77% and 92% of all relevant material recorded on MEDLINE at a price of having to sift through approximately 12.5 titles and/or abstracts to find one that refers to an article on diagnosis. This 12.5 does not seem too bad until one realizes that one deals with around 625 abstracts when 50 relevant articles are found in MEDLINE. Clinical end-users could rely on the filter with high specificity and win valuable time by reducing the number needed to read figure from 12.5 to 2.5. However, the price that must be paid is that almost one out of every two relevant articles will be missed (sensitivity of 55%). Since we know from systematic reviews that in diagnostic research in particular, there is usually great variability in study outcomes, taking the high-specificity approach can be risky. To our knowledge, no data currently show that the articles that one finds are a random selection of the available ones. This implies that a biased picture based on only half of the evidence cannot be excluded. For this reason, we do not recommend clinicians to rely on the high-specificity filter in PubMed. We found it hard to identify any group of users that may be content with a precise search that lacks good sensitivity. Among systematic reviewers there is generally consensus about the need for a sensitive search strategy. But even in reviews, precision is still very valuable since large numbers of retrieved references cannot usually be avoided with the possibility that tiredness or boredom will influence a reviewer's accuracy of study selection. In fact, the gold standard is based on hand-searches and complete articles, whereas until now the filters are assessed on their ability to identify the abstracts. Theoretically, retrieving the abstract is not a guarantee for its corresponding article to be ordered, since this depends on the reviewer's vigilance and judgment while sifting through the abstracts.

Surprisingly, the term that performed best in our search (predict*) was not evaluated as a text word by Haynes et al.[1] to build the sensitive diagnostic filter. In contrast to that study, we included commentaries, correspondences and editorials if they provided information about diagnostic tests in order to obtain valid estimates of precision. This implies that the precisions reported by Haynes et al. are likely to be somewhat overestimated.

In theory, four factors may influence a filter's reproducibility in another setting. First is the selection of journals. Second is the fact that over time the way in which abstracts are written may change (see, for example, the STARD initiative[9]). Third, over time editorial processing may change, leading to different wording in abstracts. Finally, there may be variation in the meticulousness of indexing quality in MEDLINE. Therefore, validation of any filter may range from split-sample techniques within the same (split) universe of articles to testing in other years of the same journals to other journals in other years. We found that sensitivity of our filter was stable around values of 95.2 to 98.1 to 96.1% in the 1989, 1994, and 1999 subsets, whereas precision tended to be unstable, ranging from 16.9 to 12.0 to 5%. In the 1994 and 1999 subsets, the currently available PubMed filter achieved values near its cross-validated sensitivity of 86%, namely, 96.2 and 88.5% respectively. Its precision also suffered from circumstances in which the prevalence of diagnostic studies was low (the 1999 subset had a prevalence of 61/7936 = 0.77%), yielding values of 8.2 and 4.3%, respectively. This may imply that these filters (as a diagnostic test in patients) need more extensive validation in order to characterize them better.

Table 3 ■

Description of Search Strategy Syntax for Three Commonly Used Interfaces

MEDLINE Interface	Search Syntax
Datastar	Sensitivity-and-specificity# predict$ diagnos$5 accura$
Ovid	Exp sensitivity-and-specificity or predict$ or diagnos$ or di.fs. or du.fs. or accura$
PubMed	"Sensitivity and Specificity"[MESH] OR predict* OR diagnose* OR diagnosi* OR diagnost* OR accura*
Silverplatter	Exp sensitivity-and-specificity or predict* or diagnos*or accura*

Further research should evaluate the real-world impact of the differences between our filter and the one currently implemented in PubMed in terms of time investments (cost) and consequences (missing useful papers and screening too many irrelevant ones) for clinicians and systematic reviewers alike.[10] In analogy to assessment of the impact of language restrictions on summary measures in systematic reviews,[11] the consequences of different filters or search strategies on the eventual summary measures and conclusions of diagnostic reviews could be evaluated.

Conclusion

The free-text terms identified using word frequency analysis allowed us to build and validate an alternative search filter for the detection of diagnostic studies in MEDLINE that appears to have better precision than the one currently available in PubMed while at least maintaining the latter's high sensitivity.

The authors thank Julie Glanville from the Centre for Reviews and Dissemination in York, England, for providing the word frequencies analysis.

References ■

1. Haynes RB, Wilczynski N, McKibbon KA, et al. Developing optimal search strategies for detecting clinically sound studies in MEDLINE. J Am Med Inform Assoc. 1994;1:447–58.
2. Allison JJ, Kiefe CI, Weissman NW, et al. The art and science of searching MEDLINE to answer clinical questions. Finding the right number of articles. Int J Technol Assess Health Care. 1999;15:281–96.
3. 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.
4. Cochrane Methods Working Group on Systematic Review of Screening and Diagnostic Tests. Recommended Methods. last updated on 9 February 1998. 1996. Available at <http://sm.flinders.edu.au/fusa/cochrane/>.
5. PubMed. <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi>, 2001. Accessed 3/2/02.
6. Reference Manager (10). 2002. ISI ResearchSoft. 1. Berkeley, CA 94710.
7. Boynton J, Glanville J, McDaid D, Lefebvre C. Identifying systematic reviews in MEDLINE: Developing an objective approach to search strategy design. J Inform Sci. 1998; 24:137–57.
8. Blackwell Idealist (Multiple user variant). 1-7-1994. Blackwell Software Publications, Oxford, 1994.
9. The STARD Initiative: Towards Complete and Accurate Reporting of Studies on Diagnostic Accuracy. Available at: <http://www.consort-statement.org/stardstatement.htm>. Accessed 25/02/02.
10. Jadad AR, McQuay HJ. A high-yield strategy to identify randomized controlled trials for systematic reviews. Online J Curr Clin Trials 1993; doc no 33:3973.
11. Egger M, Zellweger-Zahner T, Schneider M, et al. Language bias in randomised controlled trials published in English and German. Lancet. 1997;350:326–9.