Research Paper ■

Optimal Search Strategies for Detecting Clinically Sound Prognostic Studies in EMBASE: An Analytic Survey

Nancy L. Wilczynski, MSC, R. Brian Haynes, MD, PhD

Abstract **Background:** Clinical end users of EMBASE have a difficult time retrieving articles that are both scientifically sound and directly relevant to clinical practice. Search filters have been developed to assist end users in increasing the success of their searches. Many filters have been developed for the literature on therapy and reviews for use in MEDLINE, but little has been done for use in EMBASE with no filter development for studies of prognosis. The objective of this study was to determine how well various methodologic textwords, index terms, and their Boolean combinations retrieve methodologically sound literature on the prognosis of health disorders in EMBASE.

Methods: An analytic survey was conducted, comparing hand searches of 55 journals with retrievals from EMBASE for 4,843 candidate search terms and 8,919 combinations. All articles were rated using purpose and quality indicators, and clinically relevant prognostic articles were categorized as "pass" or "fail" according to explicit criteria for scientific merit. Candidate search strategies were run in EMBASE, the retrievals being compared with the hand search data. The sensitivity, specificity, precision, and accuracy of the search strategies were calculated.

Results: Of the 1,064 articles about prognosis, 148 (13.9%) met basic criteria for scientific merit. Combinations of search terms reached peak sensitivities of 98.7% with specificity at 50.6%. Compared with best single terms, best multiple terms increased sensitivity for sound studies by 12.2% (absolute increase), while decreasing specificity (absolute decrease 5.1%) when sensitivity was maximized. Combinations of search terms reached peak specificities of 93.4% with sensitivity at 50.7%. Compared with best single terms, best multiple terms increased specificity for sound studies by 7.1% (absolute increase), while decreasing sensitivity (absolute decrease 8.8%) when specificity was maximized.

Conclusion: Empirically derived search strategies combining indexing terms and textwords can achieve high sensitivity or specificity for retrieving sound prognostic studies from EMBASE.

■ **J Am Med Inform Assoc.** 2005;12:481–485. DOI 10.1197/jamia.M1752.

Clinicians are frequently faced with patient care questions relating to the course (prognosis) of a disease or condition. Prognostic information is also essential for planning clinical studies and health services. Clinicians and researchers can attempt to obtain answers to their patient care and planning questions in a number of ways, one of which is searching online for evidence from published investigations. The current best evidence published in health care journals is usually first widely accessible through major biomedical databases such as MEDLINE and EMBASE. Research has shown that clinicians increasingly use online access to evidence in the course of clinical care as well as for continuing education and research.[1] Research has also shown that clinicians rate the information retrieved during patient care searches as useful in answering their question.[2] However, information retrieval in these databases can be problematic. Problems arise due to the scatter of relevant articles across a broad array of journals, the very dilute concentration of high-quality, relevant studies in a very large database, and the inherent limitations of indexing, amplified by clinicians' lack of search skills.[3] EMBASE searches, for example, take place in a database containing more than nine million citations from more than 4,600 journals with between 6,000 and 8,000 citations added weekly.[4]

Researchers have developed search strategies to assist clinicians and researchers with searching, the majority of which have been developed for MEDLINE when searching for therapy, review, and diagnostic articles.[5–15] However, in addition to searching MEDLINE, clinicians may wish to search other electronic databases such as EMBASE to more comprehensively cover their topic of interest. EMBASE searching is complementary to MEDLINE searching in that EMBASE provides greater journal coverage of the European and non-English language publications as well as broader topic coverage in such areas as drug testing, toxicology, and psychiatry.[4] Fewer empirically derived search strategies have been reported for EMBASE,[16] and no work has been done in the area of prognosis.

Affiliations of the authors: Health Information Research Unit, Department of Clinical Epidemiology and Biostatistics (NLW, RBH), Department of Medicine (RBH), Michael G. DeGroote School of Medicine, McMaster University, Hamilton, Ontario, Canada.

Funded by the National Library of Medicine.

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 in the Health Information Research Unit, Department of Clinical Epidemiology and Biostatistics at McMaster University, Hamilton, Ontario, Canada.

Correspondence and reprints: R. Brian Haynes, MD, PhD, Clinical Epidemiology and Biostatistics, McMaster University, Room 2C10b, 1200 Main Street West, Hamilton, Ontario, L8N 3Z5, Canada; e-mail: <bhaynes@mcmaster.ca>.

Received for publication: 11/22/04; accepted for publication: 02/11/05.

In the early 1990s, our group at McMaster University developed search filters on a small subset of ten journals and for four types of journal articles (therapy, diagnosis, prognosis, and causation [etiology]).[17,18] This research was updated and expanded using data from 161 journals indexed in MEDLINE from the publishing year 2000.[19–22] These search strategies have been adapted for use in the Clinical Queries interface of MEDLINE (http://www.ncbi.nlm.nih.gov/entrez/query/static/clinical.html). Compared with the search strategies developed in 1991, these new strategies have been developed using a very rigorous standard. For example, the methods we applied for selecting articles were tighter, the database was much larger (161 journals compared with ten in 1991), and many more search strategies were tested, resulting in the development of search strategies that work better than the ones previously reported. We now report the extension of this research for EMBASE, including the information retrieval properties of single terms and combinations of terms for identifying methodologically sound studies on the prognosis of health disorders. To our knowledge, this is the first time that this approach has been applied to EMBASE and the first attempt of any sort to systematically retrieve methodologically sound studies of prognosis from EMBASE.

We compared the retrieval performance of methodologic search terms in EMBASE (using the Ovid interface to EMBASE) with a manual review (hand search) of each article for each issue of 55 journal titles for the year 2000. Overall, research staff hand searched 170 journal titles that were chosen based on recommendations of clinicians and librarians, Science Citation Index Impact Factors provided by the Institute for Scientific Information, and ongoing assessment of their yield of studies and reviews of scientific merit and clinical relevance for the disciplines of internal medicine, general medical practice, mental health, and general nursing practice (list of journals provided by the authors upon request). Of these 170 hand searched journals, 135 were indexed in EMBASE. Search strategies were developed using a 55 journal-subset chosen based on those journals that had the highest number of methodologically sound studies. These 55 journals were also indexed in MEDLINE and were included in the list of 161 journals that were used to develop MEDLINE search strategies.[19–22] This selection enriches the sample of target articles (those that "pass" for scientific merit), thereby improving the precision of estimates of search term performance and simplifying data processing while not biasing the estimates of the sensitivity and specificity of search terms.

We compiled an initial list of search terms, including indexing terms and textwords from clinical studies. Input was then sought from clinicians and librarians in the United States and Canada through interviews of known searchers and requests at meetings and conferences. We compiled a list of 5,385 terms of which 4,843 were unique and 3,524 returned results (list of terms tested provided by the authors upon request). Examples of the prognosis search terms tested are "inception cohort," "life expectancy," "predict," and "prognostic," all as textwords; "survival," the index term, and the index term "disease course" exploded (i.e., including all of this term's indexing subheadings).

As part of a larger study,[23] research staff with a Master's degree level of training in epidemiology and/or library science were rigorously calibrated over a 14-month period before reviewing the journals, and interrater agreement for identifying the purpose of articles was 81% beyond chance (kappa statistic, 95% confidence interval [CI] 0.79–0.84). Interrater agreement for which articles met all scientific criteria was 89% (CI 0.78–0.99) beyond chance.[23] The six research assistants then hand searched all articles in each issue of the 55 journals and applied methodologic criteria to determine whether the article was methodologically sound. The methodologic criteria applied for studies of prognosis were as follows: Inception cohort of individuals all initially free of the outcome of interest, follow-up of at least 80% of patients until the occurrence of a major study end point or to the end of the study, and analysis consistent with study design.

The proposed search strategies were treated as "diagnostic tests" for sound studies and the manual review (hand search) of the literature was treated as the gold standard. We determined the sensitivity, specificity, precision, and accuracy of each single term and combinations of terms in EMBASE using an automated process. Sensitivity for a given topic is defined as the proportion of high-quality articles for that topic that are retrieved, specificity is the proportion of low-quality articles not retrieved, precision is the proportion of retrieved articles that are of high quality, and accuracy is the proportion of all articles that are correctly classified.

Individual search terms with sensitivity >25% and specificity >75% for a given purpose category were incorporated into the development of search strategies that included two or more terms. All combinations of terms used the Boolean OR, for example, "predict.tw. OR survival.sh." For the development of multiple-term search strategies to either optimize sensitivity or specificity, we tested all two-term search strategies with sensitivity at least 75% and specificity at least 50%. For optimizing accuracy, two-term search strategies with accuracy >75% were considered for multiple-term development. In the development of prognosis search filters, 8,919 search strategies were tested.

In addition to developing search strategies using the Boolean approach described above, we also evaluated the potential for improving performance using logistic regression. Two approaches were taken. First, we took the top performing Boolean search strategies and ORed additional terms to these base strategies using stepwise logistic regression. The level of significance for entering and removing search terms from the model was 0.05. Adding terms to the model stopped when the increase in the area under the receiver operating characteristic curve was <1%. Second, we developed search strategies from scratch with stepwise logistic regression using these same cutoff values. Both logistic regression approaches were compared with the Boolean approach to search strategy development when developing strategies for treatment articles and prognostic articles for MEDLINE. Treatment and prognosis were chosen because they represented the best and the worst cases for MEDLINE search strategy performance. For both purpose categories, the logistic regression approaches to developing search strategies did not improve performance compared with search strategies developed using the Boolean approach described above. We also found that when strategies were developed in 60% of the database and validated in the remaining 40%, there were no statistical differences in performance. Thus, for subsequent purpose categories and databases, including EMBASE, the Boolean approach was

Table 1 ■ Single Term with the Best Sensitivity (Keeping Specificity ≥50%), Best Specificity (Keeping Sensitivity ≥50%), and Best Optimization of Sensitivity and Specificity (Based on the Lowest Possible Absolute Difference between Sensitivity and Specificity) for Detecting Studies of Prognosis in EMBASE in 2000

Search Term Ovid Search*	Sensitivity (%) (95% CI) (*n* = 148)	Specificity (%) (95% CI) (*n* = 27,621)	Precision (%) (95% CI)†	Accuracy (%) (95% CI) (*n* = 27,769)
Best sensitivity				
exp general aspects of disease	86.5 (81.0–92.0)	55.7 (55.1–56.3)	1.0 (0.9–1.2)	55.7 (55.3–56.5)
Best specificity				
exp disease course	59.5 (51.6–67.4)	86.3 (86.0–86.8)	2.3 (1.8–2.8)	86.2 (85.8–86.6)
Best optimization of sensitivity and specificity				
exp physical disease by body function	51.4 (43.3–59.4)	58.7 (58.1–59.3)	0.7 (0.5–0.8)	58.7 (58.1–59.3)

CI = confidence interval; exp = exploded subject heading.
*The search strategy is reported using Ovid's search engine syntax for EMBASE.
†Denominator varies by row.

Table 2 ■ Combination of Terms with the Best Sensitivity (Keeping Specificity ≥50%), Best Specificity (Keeping Sensitivity ≥50%), and Best Optimization of Sensitivity and Specificity (Based on abs[sensitivity – specificity]<1%) for Detecting Studies of Prognosis in EMBASE in 2000

Search Strategy Ovid Search*	Sensitivity (%) (95% CI) (*n* = 148)	Specificity (%) (95% CI) (*n* = 27,621)	Precision (%) (95% CI)†	Accuracy (%) (95% CI) (*n* = 27,769)
Best sensitivity				
exp disease course OR risk:.mp. OR diagnos:.mp. OR follow-up.mp. OR ep.fs. OR outcome.tw.	98.7 (96.8–100.0)	50.6 (50.0–51.2)	1.1 (0.9–1.2)	50.9 (50.3–51.4)
Best specificity				
prognos:.tw. OR survival.tw.	50.7 (42.6–58.7)	93.4 (93.1–93.7)	3.9 (3.1–4.8)	93.2 (92.9–93.5)
Small drop in specificity with a substantive gain in sensitivity				
prognos:.tw. OR surviv:.tw.	58.1 (50.2–66.1)	92.5 (92.2–92.8)	4.0 (3.2–4.8)	92.4 (92.0–92.7)
Best optimization of sensitivity and specificity				
follow-up.mp. OR prognos:.tw. OR ep.fs.	80.4 (74.0–86.8)	79.9 (79.4–80.4)	2.1 (1.7–2.5)	79.9 (79.4–80.4)

CI = confidence interval; exp = exploded subject heading; : = truncation; mp = multiple posting—term appears in title, abstract, or subject heading; ep = epidemiology; fs = floating subheading; tw = textword (word or phrase appears in title or abstract).
*Search strategies are reported using Ovid's search engine syntax for EMBASE.
†Denominator varies by row.

used for search strategy development and search strategies were developed using all records in the database.

Results

Indexing information was downloaded from EMBASE for 27,769 articles (excluding duplicates) from the 55 journals hand searched. Of these, 1,064 were classified as prognosis, of which 148 (13.9%) were methodologically sound. Search strategies were developed using all 27,769 articles. Thus, the strategies were tested for their ability to retrieve articles about high-quality prognosis studies from all other articles, including both low-quality prognosis studies and all nonprognosis studies.

Table 1 shows the best single term for high sensitivity, high specificity, and best balance of sensitivity and specificity. The single term "exp general aspects of disease" produced the best sensitivity of 86.5% while keeping specificity at 55.7%. Specificity was maximized at 86.3% using the single term "exp disease course," but this was achieved at the expense of sensitivity (59.5%). The single term "exp physical disease by body function" produced the optimal balance between sensitivity (51.4%) and specificity (58.7%).

Combination of terms with the best results for sensitivity, specificity, and optimization of sensitivity and specificity are shown in Table 2. The six-term search strategy "exp disease course OR risk:.mp. OR diagnos:.mp. OR follow-up.mp. OR ep.fs. OR outcome.tw." achieved a sensitivity of 98.7% with a specificity at 50.6%. The two-term strategy "prognos:.tw. OR survival.tw." had the highest specificity at 93.4%, outperforming all three-term combinations. A three-term combination "follow-up.mp. OR prognos:.tw. OR ep.

fs.'' resulted in the optimization strategy achieving approximately 80% for both sensitivity and specificity (Table 2).

A slight modification to the above-noted most specific search strategy led to an attractive trade-off in sensitivity and specificity (Table 2). By replacing ''survival.tw.'' with ''surviv:.tw.'' in the most specific search strategy (''prognos:.tw. OR survival.tw.''), sensitivity increased (50.7% to 58.1%) at the price of a small decrease in specificity (93.4% to 92.5%).

Discussion

Our study documents search strategies that can help discriminate relevant, high-quality studies from lower quality studies of the prognosis of health disorders and articles that are not about prognosis. Those interested in all articles on prognosis, for example, those conducting systematic reviews, will be best served by the most sensitive search. Those with little time on their hands who are looking for a few good articles on prognosis, for example, clinicians looking for answers to patient care questions, will likely be best served by the most specific strategies. The strategies that optimized sensitivity and specificity while minimizing the difference between the two provide the best separation of target citations from undesired citations but do so without regard for whether sensitivity and specificity are affected.

All search strategies had low precision, which was expected because of the low proportion of relevant studies about prognosis in a very large, multipurpose database. This means that searchers will continue to need to spend time discarding irrelevant retrievals. Low values for precision, while of concern, should not be overinterpreted because we did not limit the searches by clinical content terms, as would be the case in clinical patient care searches. Precision might be enhanced by combining search strategies in these tables with content specific terms using the Boolean ''AND'' and/or by combining search strategies with methodologic terms using the Boolean ''AND NOT.'' Additionally, conducting searches in journal subsets might enhance precision. We are currently testing these types of more sophisticated strategies as the next phase of our project.

Comparing the prognostic search strategies developed for EMBASE with those that we developed for MEDLINE,[22] we find that top-performing single terms were all index terms in both EMBASE (Table 1) and MEDLINE (''exp epidemiologic studies'' was the top performer for sensitivity, specificity, and optimization) but that these terms are uniquely supported by the database in question (i.e., the index terms shown for EMBASE are not found in MEDLINE and vice versa). Additionally, we find that for multiple-term strategies, a mix of index words and textwords are required in both databases, but the combination of terms that perform best is very different. The only textword that was a top performer in both databases was ''prognos:.tw.'' Although there are many differences between EMBASE and MEDLINE, some basic similarities are apparent as just described.

Other methods of improving bibliographic retrieval exist, including statistical approaches (such as multiple logistic regression) and machine learning models. Logistic regression did not improve the retrieval of articles in MEDLINE concerning treatment and prognosis in our database.[21] Aphinyanaphongs and colleagues[24] have reported machine learning-based enhancements of retrieval of clinical studies from MEDLINE in comparison with our previously published search strategies.[17] We would welcome head-to-head comparisons of machine learning and other approaches with our new ''brute force'' strategies in MEDLINE, EMBASE, or other bibliographic databases.

In conclusion, our study has shown that selected combinations of indexing terms and textwords can achieve high sensitivity or specificity in retrieving prognosis studies cited in EMBASE.

References ■

1. 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.
2. Magrabi F, Westbrook JI, Coiera EW, Gosling AS. Clinicians' assessments of the usefulness of online evidence to answer clinical questions. Medinfo. 2004;11(pt 1):297–300.
3. 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.
4. STN Database Summary Sheet. EMBASE. Available from: http://www.cas.org/ONLINE/DBSS/embasess.html Accessed August 6, 2004.
5. Robinson KA, Dickersin K. Development of a highly sensitive search strategy for the retrieval of reports of controlled trials using PubMed. Int J Epidemiol. 2002;31:150–3.
6. Nwosu CR, Khan KS, Chien PF. A two-term MEDLINE search strategy for identifying randomized trials in obstetrics and gynecology. Obstet Gynecol. 1998;91:618–22.
7. Marson AG, Chadwick DW. How easy are randomized controlled trials in epilepsy to find on Medline? The sensitivity and precision of two Medline searches. Epilepsia. 1996;37:377–80.
8. Adams CE, Power A, Frederick K, LeFebvre C. An investigation of the adequacy of MEDLINE searches for randomized controlled trials (RCTs) of the effects of mental health care. Psychol Med. 1994;24:741–8.
9. Dumbrigue HB, Esquivel JF, Jones JS. Assessment of MEDLINE search strategies for randomized controlled trials in prosthodontics. J Prosthodont. 2000;9:8–13.
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.
11. Shojania KG, Bero LA. Taking advantage of the explosion of systematic reviews: an efficient MEDLINE search strategy. Eff Clin Pract. 2001;4:157–62.
12. 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.
13. 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.
14. 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.
15. Vincent S, Greenley S, Beaven O. Clinical Evidence diagnosis: Developing a sensitive search strategy to retrieve diagnostic studies on deep vein thrombosis: a pragmatic approach. Health Info Libr J. 2003;20:150–9.
16. Bachmann LM, Estermann P, Kronenberg C, ter Riet G. Identifying diagnostic accuracy studies in EMBASE. J Med Libr Assoc. 2003;91:341–6.

17. 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.
18. 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.
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, Ramkissoonsingh R; Hedges Team. Developing optimal search strategies for detecting sound clinical prediction studies in MEDLINE. AMIA Annu Symp Proc. 2003;728–32.
21. Haynes RB, Wilczynski NL. Optimal search strategies for retrieving scientifically strong studies of diagnosis from MEDLINE: an analytical survey. BMJ. 2004;1:328:1040–2.
22. Wilczynski NL, Haynes RB; Hedges Team. Developing optimal search strategies for detecting clinically sound prognostic studies in MEDLINE: an analytic survey. BMC Med. 2004;2:23.
23. 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; 10:390–3.
24. Aphinyanaphongs Y, Tsamardinos I, Statnikov A, Hardin D, Aliferis CF. Text categorization models for high quality article retrieval in internal medicine. J Am Med Inform Assoc. 2005; 12:207–16.