Assessment of Methodologic Search Filters in MEDLINE

NL Wilczynski, CJ Walker, KA McKibbon, RB Haynes
Health Information Research Unit,
Dept. of Clinical Epidemiology & Biostatistics,
McMaster University, 1200 Main St W, Hamilton, Ont, Canada L8N 3Z5
(416)525-9140 x2311, FAX 416-546-0401, E-MAIL WILCZYNS@McMASTER

ABSTRACT

Objective: To determine the retrieval characteristics of methodologic textwords and MeSH terms in MEDLINE for identifying methodologically sound studies on the etiology, prognosis, diagnosis, and prevention and treatment of disorders in general adult medicine.
Design: Comparison of methodologic search terms and phrases for the retrieval of citations in MEDLINE with a manual hand search of the literature (the gold standard) for 10 internal and general medicine journals for 1986 and 1991.
Measures: Sensitivity (proportion of methodologically sound and correct topic studies retrieved) and specificity (proportion of unsound or wrong topic articles not retrieved) of the search strategies.
Results: The individual terms yielding the best sensitivity for 1991 by purpose category were: risk (tw) for etiology; exp cohort studies for prognosis; sensitivity (tw) for diagnosis; and clinical trial (pt) for treatment. The corresponding terms for 1986 were: risk (tw) for etiology; prognos: (tw) for prognosis; sensitivity (tw) for diagnosis; and random: (tw) for treatment.
Conclusions: The performance of methodologic MeSH terms and textwords varied greatly in MEDLINE and changed from 1986 to 1991. More complex search strategies may be required to optimize retrieval.

INTRODUCTION

It is important for clinical end users of MEDLINE to be able to retrieve articles that are both scientifically sound and directly relevant to clinical practice. The use of "methodologic search filters" (such as 'random allocation' for sound studies of medical interventions) has been advocated in order to improve the accuracy of searching for such studies. [1]. However, the retrieval performance of such terms on search recall and precision has not been tested. The recall and precision of methodologic search terms may be enhanced by combinations of terms and by mixing textwords and Medical Subject Headings (MeSH) in search strategies. The purpose of this study was to test individual methodologic MeSH terms and textwords in common use, and permutations and combinations of these MeSH terms and textwords for identifying methodologically sound studies on the etiology, prognosis, diagnosis, and prevention and treatment of disorders in general adult medicine. In this paper, we report on the information retrieval properties of single terms. Our results are of most interest to clinicians doing their own searches for clinically relevant and valid studies and for librarians involved in assisting clinicians to construct their own searches.

METHODS

To evaluate MEDLINE strategies designed to retrieve methodologically rigorous studies, search terms and textwords related to research design features were treated as "diagnostic tests" or screening procedures for the detection of relevant citations. Borrowing from the concepts of diagnostic test evaluation, the sensitivity and specificity of MEDLINE searches were determined. The search strategies were designed to detect methodologically sound studies of human adult care among all original and review articles appearing in selected internal medicine and general medical journals. The yield of MEDLINE searches was determined by comparison with manual hand searches of the journals, the gold standard. Thus, sensitivity of the MEDLINE search strategies was calculated as the proportion of correctly detected citations with relevant content and sound study methods among all relevant citations as defined by the hand search. This is equivalent to the library term 'recall'. Specificity was the proportion of unsound studies and irrelevant articles excluded by the search strategy. This differs from precision which is the proportion of all articles retrieved by a search strategy that are sound and relevant.

Hand Search of the Literature

For the years 1986 and 1991, three research assistants hand searched 10 journals, the same 10 in each year, for methodologically sound articles on the etiology, prognosis, diagnosis, and prevention and treatment of human adult disease. The ten journals searched were *American Journal of Medicine, Annals of Internal Medicine, Archives of Internal*

0195-4210/92/$5.00 © 1994 AMIA, Inc.

Medicine, BMJ (British Medical Journal in 1986), *Circulation, Diabetes Care, Journal of Internal Medicine (Acta Medica Scandinavica* in 1986), *Journal of the American Medical Association, The Lancet,* and *New England Journal of Medicine,* including supplements. These journals were selected to provide a broad range of publications, including both internal and general medical journals, and both American and European authors.

Articles were classified for 'format', 'interest', 'purpose' and 'methodologic rigor'. Format categories included 'original study', 'review', 'general article', 'conference report', 'decision analysis', and 'case report'. Articles with more than one format were classified for all that applied. For the purpose of this investigation an original study was defined as any full text article in which the investigators had made first hand observations from life, case records or other sources of information. A review was any full text article that was bannered review, had review in the title or in a section heading, or indicated in the text that the intention was to review or summarize the literature on a topic. A general article was a general or philosophical discussion of a topic without original first-hand observation or a statement that the purpose was to review or appraise a body of knowledge, including unbannered news items, unbannered editorials, position and opinion papers, musings and psychosocial observations. A conference report was defined as such by the journal but was classified as an original or review article when meeting those criteria. A decision analysis was defined as the breaking down of the management of patients into component parts, defining routes of management and consequences of management based on alternatives, for the purpose of defining optimal methods of management. A case report was defined as an original study involving less than 10 subjects. Journal items excluded from classification included bannered letters to the editor, book reviews, announcements, policy watch, editorials, commentaries, brief clinical observations, correspondence, news, obituaries, postgraduate and continuing education forums, and notices.

To be considered of interest to the medical care of human adults the study had to be concerned with the understanding and management of clinical problems with clinical endpoints and recommendations for applications in human subjects, at least 50% of whom were ≥ 18 years of age at study entry. All format categories were classified for interest.

Articles classified as original studies, reviews, or case reports and of interest were classified for purpose. Articles could have more than one purpose and were classified for all that applied. Articles were classified as 'etiology' when the content pertained directly to causation of a disease or condition; as 'prognosis' when the content pertained directly to the prediction of the clinical course or the natural history of a disease with the disease existing at the beginning of the study; as 'diagnosis' when the content pertained directly to the evaluation of a disease process, usually through comparing methods of arriving at a diagnosis; as 'treatment or prevention' when the content pertained directly to therapy, prevention or rehabilitation; and as 'something else' when the purpose of the study was something other than the above.

Studies in each purpose category were evaluated for methodologic rigor and deemed methodologically sound if they fulfilled one criterion specific to their purpose. These criteria were chosen according to critical appraisal criteria for applied research [2]. Original studies of interest and classified as etiology were considered methodologically sound if there was a formal control group achieved when one of the following was evident: there was random or quasi-random allocation of participants to treatment and control groups; or the study was a non-randomized, concurrent control trial, a cohort analytic study with case-by-case matching or statistical adjustment to create comparable groups, or a case control study. Original studies of interest and classified as prognosis were considered methodologically sound if there was a cohort of subjects all having the disease in question at baseline without the outcome of interest. Original studies of interest and classified as diagnosis were considered methodologically sound if there was provision of sufficient data to calculate the sensitivity and specificity of the test or likelihood ratios based on subjects who had all been tested on both the test and diagnostic standard. Original studies of interest and classified as treatment or prevention were considered methodologically sound if there was random or quasi-random allocation of participants to treatment and control groups. Review articles of interest could be assigned to one or more purpose categories but were considered methodologically sound if there was an identifiable reproducible description of the methods for conducting the literature review.

Inter-rater reliability was assessed for the classification of articles for format, interest, purpose and methods. In all cases the degree of agreement beyond chance was assessed by the Kappa statistic and was greater than 0.80.

The sample size required to detect a 20% improvement in sensitivity for the comparison of one MEDLINE search strategy with another on the same topic was 73 methodologically sound studies in each of the purpose categories for each of the years 1986 and 1991 (type 1 error of 5%, one-sided, and a type 2 error rate of 20%).

Collecting terms and MEDLINE searches

To collect a comprehensive list of terms and phrases that were to be tested against the hand search, input was sought from clinicians and librarians in the United States and Canada. Known searchers were interviewed, requests were placed on several bulletin boards and in national publications, input was sought at meetings and conferences, and requests were sent to the National Library of Medicine and Canada Institute for Scientific and Technical Information. Individuals were asked what terms or phrases they used when searching for etiology, prognosis, diagnosis, and prevention and treatment articles and related review articles. MeSH, Publication Types, Check Tags, and subheadings that indexers at the National Library of Medicine used for indexing articles were solicited. Methodologic words or phrases (textwords) that appeared often in titles and abstracts of articles were also selected.

From the submissions a list was compiled of the terms and phrases by purpose category, duplicates were eliminated, and terms that were inaccurate or redundant were discarded. Some of the terms and phrases were different for the 2 years as publication types were introduced in 1991 and some of the corresponding terms changed definitions. Also, some terms retrieved 0 citations for the 10 journals in 1986 and were discarded for this year.

DATA COLLECTION

Hand ratings of the 10 journals for 1991 and 1986 were recorded on data collection forms, and the bibliographic information, including the 8-digit unique identifier, for the articles in those journals was captured from MEDLINE. Each journal title was searched in MEDLINE for 1991 and 1986 and the publication types 'editorial,' 'comment,' 'letter' and 'news' were eliminated from the search using the Boolean AND NOT operator.

The terms to be tested were searched in MEDLINE and the unique identifiers were captured. Search strategies were made up of only methodologic terms.

TESTING STRATEGIES

All methods terms were tested, both individually and in combination, in terms of their sensitivity, specificity and precision. For 1991 there were 25 etiology terms, 27 prognosis terms, 22 diagnosis terms and 20 treatment terms. For 1986 there were 18 etiology terms, 21 prognosis terms, 22 diagnosis terms and 14 treatment terms (see Appendix for list of terms).

RESULTS

The hand search defined which articles were of relevant content (original, review and case report articles) and methodologically sound. The total number of original, review, and case report articles and the breakdown by 'format' category and by 'methodologic rigor' are presented in Table 1.

Table 1
Number of Original, Review and Case Report Articles and Breakdown by 'Format' and 'Methodologic Rigor'

	1986	1991
Total No. of Original, Review & Case Report Articles	3682	3495
No. Classified as Etiology (ET)	531	523
No. of ET Methodologically Sound	155	201
No. Classified as Prognosis (PR)	149	205
No. PR Methodologically Sound	106	133
No. Classified as Diagnosis (DI)	426	412
No. DI Methodologically Sound	92	111
No. Classified as Treatment (TR)	936	879
No. TR Methodologically Sound	270	281
No. Classified as Review Articles (R)	337	543
No. R Methodologically Sound	4	47

For 1991, the terms with ≥ 50% sensitivity are presented in Table 2. Terms with < 10% sensitivity are marked with an asterisk (*) in the Appendix.

For 1986, the terms with ≥ 50% sensitivity are presented in Table 3. Terms with < 10% sensitivity are marked with a hat (^) in the Appendix.

Table 2
Single Terms - Sensitivity ≥ 50%

1991			
CATEGORY	STRATEGY	SENSITIVITY	SPECIFICITY
Etiology	Risk (tw)	0.67	0.79
	Exp Risk	0.58	0.89
	Exp Causality	0.51	0.90
	Risk Factors	0.51	0.90
Prognosis	Exp Cohort Studies	0.60	0.80
	Exp Longitudinal Studies	0.56	0.83
	Prognos: (tw)	0.52	0.96
Diagnosis	Sensitivity (tw)	0.57	0.97
	Specificity (tw)	0.54	0.98
	Exp Sensitivity and Specificity	0.50	0.98
Treatment	Clinical Trial (pt)	0.93	0.92
	Random: (tw)	0.89	0.92
	Randomized Controlled Trial (pt)	0.87	0.97

Table 3
Single Terms - Sensitivity ≥ 50%*

1986			
CATEGORY	STRATEGY	SENSITIVITY	SPECIFICITY
Etiology	Risk (tw)	0.61	0.89
Prognosis	Prognos: (tw)	0.56	0.97
	Exp Longitudinal Studies	0.56	0.89
	Exp Cohort Studies	0.56	0.89
	Prognosis	0.50	0.97
Treatment	Random: (tw)	0.82	0.95
	Exp Research Design	0.79	0.96
	Random Allocation	0.70	0.97
	Clinical Trials	0.70	0.96
	Clinical Trial (pt)	0.70	0.96

* The highest sensitivity for diagnosis was 0.44 (specificity 0.98) for sensitivity (tw).

DISCUSSION

The results of this study show that the performance of individual methodologic MeSH terms and textwords in MEDLINE varied greatly when attempting to retrieve methodologically sound studies on the etiology, prognosis, diagnosis, and prevention and treatment of disorders in general adult medicine. Many terms yielded a sensitivity < 10% and were therefore of no use in MEDLINE search strategies. Other terms in each purpose category performed much better, however.

The term that yielded the highest sensitivity for treatment in 1991 was 'clinical trial (pt)' (93%). This is much higher than reported in previous studies. Poynard and Conn [3], Dickersin and coworkers [4], and Kirpalani and coworkers [5] found recall rates of 51%, 29%, and 53%, respectively for randomized trials on selected topics (liver disease, neonatal hyperbilirubinemia, and care of newborn infants, respectively). These studies were conducted in or prior to 1985. This may account for the difference in the results as we noted a large improvement when comparing the sensitivities between 1986 and 1991; for example, the sensitivity for 'clinical trial (pt)' was 70% in 1986 and 93% in 1991.

The search filters presented here can aid searchers, particularly clinicians who are inexperienced in constructing complex searches, to retrieve studies that meet at least one major criterion for scientific merit for applied health care research while filtering out studies with weaker designs. Such filters are bound to retrieve some 'false positive' articles and miss others that should be retrieved. Retrieved articles must be further evaluated to determine their methodologic soundness and clinical applicability. 'False negative' articles can only be retrieved by hand searching journals or other labor-intensive means.

Other possible quality filters such as ordering journals by impact factors and citations exist but we do not know how this method compares with our search filters. However, even among the best journals only a small proportion of articles meet the quality criteria we used.

One limitation of this study was that only priority journals were included in the search. Also, only the abstracts of citations could be searched for textword inclusion. However, one of the strengths of this study was the highly reproducible classification of articles in the manual hand searches which served as the gold standard.

The results of this study showed that treatment terms performed very well in retrieving articles that were methodologically sound. The performance in the other purpose categories were less good but it may be possible to improve performance with permutations and combinations of MeSH terms and textwords. Unfortunately, there are many thousand combinations and it will take some time to work through them in the next phase of our research.

Until research on the combination of terms is completed we recommend that the term with the highest sensitivity within each purpose category be used in the MEDLINE search (for example, 'clinical trial (pt)' in searches on treatment and prevention). If the yield of articles is not satisfactory with the lead term additional terms should be 'ORed' to increase the yield. In back file searches the most appropriate term may differ and the search should be modified appropriately.

Appendix

Etiology

Indexing terms

- exp case control studies^
- case control studies (1991 only)
- retrospective studies*^
- exp cohort studies
- cohort studies (1991 only)
- exp longitudinal studies
- longitudinal studies*^
- follow-up studies^
- prospective studies
- cross-sectional studies*^
- exp causality (1991 only)
- causality (1991 only)*
- risk factors (1991 only)
- exp risk
- risk*
- logistic models (1991 only)*
- odds ratio (1991 only)*

textwords

- cohort^
- risk
- etiol: or aetiol: (the colon indicates truncation)*^
- odds and ratio:^
- causation or causal:*
- relative and risk
- case and control:
- case and comparison*^

Prognosis

Indexing terms

- exp cohort studies
- cohort studies (1991 only)*
- exp longitudinal studies
- longitudinal studies*^
- follow-up studies
- prospective studies
- prognosis
- exp morbidity^
- morbidity*^
- incidence (1991 only)
- exp mortality^
- mortality*^
- cause of death (1991 only)*
- infant mortality*^
- maternal mortality*^
- survival rate (1991 only)
- survival analysis (1991 only)

textwords

- natural and history*
- prognos:
- inception and cohort (1991 only)*
- clinical and course^
- predict:
- outcome:
- clinical and consequence:*^
- prognostic factor:
- morbidity*^
- course

Diagnosis

Indexing terms

- exp sensitivity and specificity^
- sensitivity and specificity^
- predictive value of tests^
- ROC curve (1991 only)*
- exp diagnostic errors*^
- diagnostic errors*^
- false positive reactions*^
- false negative reactions*^
- diagnosis, differential*^

textwords

- sensitivity
- specificity
- predictive and value:
- post and test and probabilit: (1986 only)^
- post and test and likelihood (1986 only)^
- likelihood and ratio:*^
- false and rate*^
- false and positive*
- false and negative*^
- receiver and operat: and characteristic*^
- roc*^
- independent and comparison*^
- mask: and comparison (1991 only)*
- blind: and comparison*^
- gold and standard*^

Treatment

Indexing terms

- exp research design
- research design*^
- double-blind method
- random allocation*
- exp clinical trials (1991 only)*
- clinical trials*
- multicenter studies (1991 only)*
- randomized controlled trials (1991 only)*
- clinical trial (pt)
- multicenter study (pt) (1991 only)
- randomized controlled trial (pt) (1991 only)
- comparative study
- single-blind method (1991 only)*
- placebos*^

textwords

- random:
- placebo:
- double and blind:
- mask:*^
- single and blind:*^
- controlled and trial:

Review Articles

Indexing terms

- review (pt)
- review literature (1991 only)
- review of reported cases (1991 only)
- review, academic (1991 only)
- review, multicase (1991 only)
- review, tutorial (1991 only)
- meta analysis (1991 only)

textwords

- MEDLINE
- meta and analysis
- overview
- review
- all review

* Terms with < 10% sensitivity in 1991.
^ Terms with < 10% sensitivity in 1986.

ACKNOWLEDGEMENTS

The study was supported by the Ontario Ministry of Health and the National Library of Medicine (R01 LM04696-03).

References

[1]. 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.

[2]. Sackett DL, Haynes RB, Guyatt GH, Tugwell P. Clinical epidemiology: a basic science for clinical medicine. Second Edition. Little, Brown and Company, Boston, 1991.

[3]. Poynard T, Conn H. The retrieval of randomized clinical trials in liver disease from the medical literature. A comparison of MEDLARS use and manual methods. Controlled Clin Trials 1985;6:271-9.

[4]. Dickersin K, Hewitt P, Mutch L, Chalmers I, Chalmers TC. Perusing the literature: comparison of MEDLINE searching with a perinatal trials database. Controlled Clin Trials 1985;6:306-17.

[5]. Kirpalani H, Schmidt B, McKibbon KA, Haynes RB, Sinclair JC. Searching MEDLINE for randomized clinical trials involving care of the newborn. Pediatrics 1989;83:543-6.