Research Paper ■

Developing Optimal Search Strategies for Detecting Clinically Sound Studies in MEDLINE

R. BRIAN HAYNES, MD, PHD, NANCY WILCZYNSKI, MSC, K. ANN MCKIBBON, MLS, CYNTHIA J. WALKER, MLS, JOHN C. SINCLAIR, MD

Abstract **Objective:** To develop optimal MEDLINE search strategies for retrieving sound clinical studies of the etiology, prognosis, diagnosis, prevention, or treatment of disorders in adult general medicine.

Design: Analytic survey of operating characteristics of search strategies developed by computerized combinations of terms selected to detect studies meeting basic methodologic criteria for direct clinical use in adult general medicine.

Measures: The sensitivities, specificities, precision, and accuracy of 134,264 unique combinations of search terms were determined by comparison with a manual review of all articles (the "gold standard") in ten internal medicine and general medicine journals for 1986 and 1991.

Results: Less than half of the studies of the topics of interest met basic criteria for scientific merit for testing clinical applications. Combinations of search terms reached peak sensitivities of 82% for sound studies of etiology, 92% for prognosis, 92% for diagnosis, and 99% for therapy in 1991. Compared with the best single terms, multiple terms increased sensitivity for sound studies by over 30% (absolute increase), but with some loss of specificity when sensitivity was maximized. For 1986, combinations reached peak sensitivities of 72% for etiology, 95% for prognosis, 86% for diagnosis, and 98% for therapy. When search terms were combined to maximize specificity, over 93% specificity was achieved for all purpose categories in both years. Compared with individual terms, combined terms achieved near-perfect specificity that was maintained with modest increases in sensitivity in all purpose categories except therapy. Increases in accuracy were achieved by combining terms for all purpose categories, with peak accuracies reaching over 90% for therapy in 1986 and 1991.

Conclusions: The retrieval of studies of important clinical topics cited in MEDLINE can be substantially enhanced by selected combinations of indexing terms and textwords.

■ **J Am Med Informatics Assoc.** 1994;1:447–458.

Clinical end-user searching of MEDLINE has risen dramatically during the past five years,[1] spurred by the development of user-friendly software, a proliferation of online and compact disc formats, falling user charges, and advertising directed at clinicians. MEDLINE, however, is a general-purpose biomedical research literature database, with only a small fraction of articles reporting evidence that can be applied directly to clinical practice. The large number of postings in MEDLINE (several million), the low prevalence of clinically applicable studies, the well-documented limitations of indexing and retrieval in

Affiliation of the authors: Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario, Canada.

Supported by The Ontario Ministry of Health (grant no. 02651) and the United States National Library of Medicine (grant no. R01 LM 04696). Dr. Haynes is supported in part by a National Health Scientist Award from the National Health Research and Development Program of Canada.

Correspondence and reprints: R. Brian Haynes, MD, PhD, McMaster University Medical Centre, Room 3H7, 1200 Main Street West, Hamilton, Ontario L8N 3Z5, Canada. e-mail: BHAYNES@MCMASTER.CA

Received for publication: 5/12/94; accepted for publication: 7/01/94.

Table 1 ▪

Formula for Calculating the Sensitivities, Specificities, Precision, and Accuracy of MEDLINE Searches for Detecting Sound Clinical Studies

		Manual Review	
		Meets Criteria	Does Not Meet Criteria
Search Terms	Detected	a	b
	Not detected	c	d
		a + c	b + d

Sensitivity = a/(a + c).
Specificity = d/(b + d).
Precision = a/(a + b + articles of other formats that are detected).
Accuracy = (a + d)/(a + b + c + d).*
*(a + b + c + d) = all original, review, and case reports as defined by the manual review of the literature.

MEDLINE from its inception,[2] and the imprecise search skills of clinical end users[3] all contribute to difficulties in using MEDLINE to seek answers to clinical questions. The problems lead to both missing sound studies in searching (low sensitivity) and retrieving many citations of studies that are not sound (low specificity and low precision).

A potential method for improving the detection of studies of high quality for clinical practice from MEDLINE is to include search terms that select studies that are at the most advanced stages of testing for clinical application. There are relatively few study designs for these final stages of testing. For example, the randomized controlled trial is widely accepted as the standard for testing a treatment or preventive procedure.[4] Similarly, an inception cohort is required for studies of prognosis, an independent "gold standard" is required for validation of a diagnostic test, and a valid comparison group is required for studies of etiology or causation. MEDLINE indexers have had "random allocation" and "randomized controlled trials" as indexing terms for some time, recently strengthened by "publication types (pt)." For diagnostic test evaluation, "sensitivity" and "specificity," "predictive value of tests," and "receiver operating characteristic (ROC) curve" are relatively recent additions to the Medical Subject Headings (MeSH) vocabulary. For prognosis, "explode cohort studies" is perhaps the closest equivalent to "inception cohort." For etiology, there are many research design options and the indexing is similarly varied, with terms such as "risk factors" and "causality" that may be applied to the content of studies regardless of the studies' methodologies or quality. Authors, however, may use more exact methodologic terms in their titles and abstracts, and these terms [textwords (tw)] may then provide another opportunity for retrieval.

We have advocated using methodologic search filters to improve the retrieval of studies of higher quality for clinical practice[5] and sought in this study to develop better methodologic search filters and to verify their validity. The information retrieval properties of combinations of terms are reported. The retrieval properties of individual terms were published previously.[6] The results of this study would be of most interest to clinicians doing their own searches for clinically relevant and valid studies and to librarians involved in assisting clinicians to construct their own searches.

Methods

The study compared the retrieval performance of methodologic search terms and phrases in MEDLINE with a manual review of each article for each issue of ten internal medicine and general medicine journals for the two years 1986 and 1991. To evaluate MEDLINE strategies designed to retrieve studies meeting basic methodologic criteria for clinical practice, MeSH terms and textwords related to research design features were run as search strategies. These search strategies were treated as diagnostic tests for sound studies and the manual review of the literature was treated as the "gold standard." Borrowing from the concepts of diagnostic test evaluation and library science, the sensitivities, specificities, accuracy, and precision of MEDLINE searches were determined as shown in Table 1. For example, the sensitivity of each MEDLINE search strategy was calculated as the proportion of relevant, sound citations detected by that strategy.

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 studies meeting the methodologic criterion in each of the purpose categories for each of the years 1986 and 1991 (a type 1 error rate of 5%, one-sided, and a type 2 error rate of 20%).

Figure 1 illustrates the steps involved in the data collection and analysis stages, as detailed below.

Manual Review of the Literature

For the years 1986 and 1991, three research assistants assessed ten journals, the same ten in each year, for articles meeting basic methodologic criteria concerning the etiology, prognosis, diagnosis, prevention, and treatment of disease of human adults. The ten journals searched were *American Journal of Medicine,* *Annals of Internal Medicine, Archives of Internal 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, Lancet,* and *New England Journal of Medicine,* including supplements. These journals were selected on the basis of high-impact factors and immediacy indexes[7] and in order to provide a range of publications, including both internal medicine and general medicine journals, and both American and European authors.

Articles were classified for format, interest, purpose, and methodologic rigor. The format categories and their corresponding definitions are shown in Table 2. Items excluded from classification included bannered letters to the editor, book reviews, announcements, policy watch, editorials, brief clinical ob-

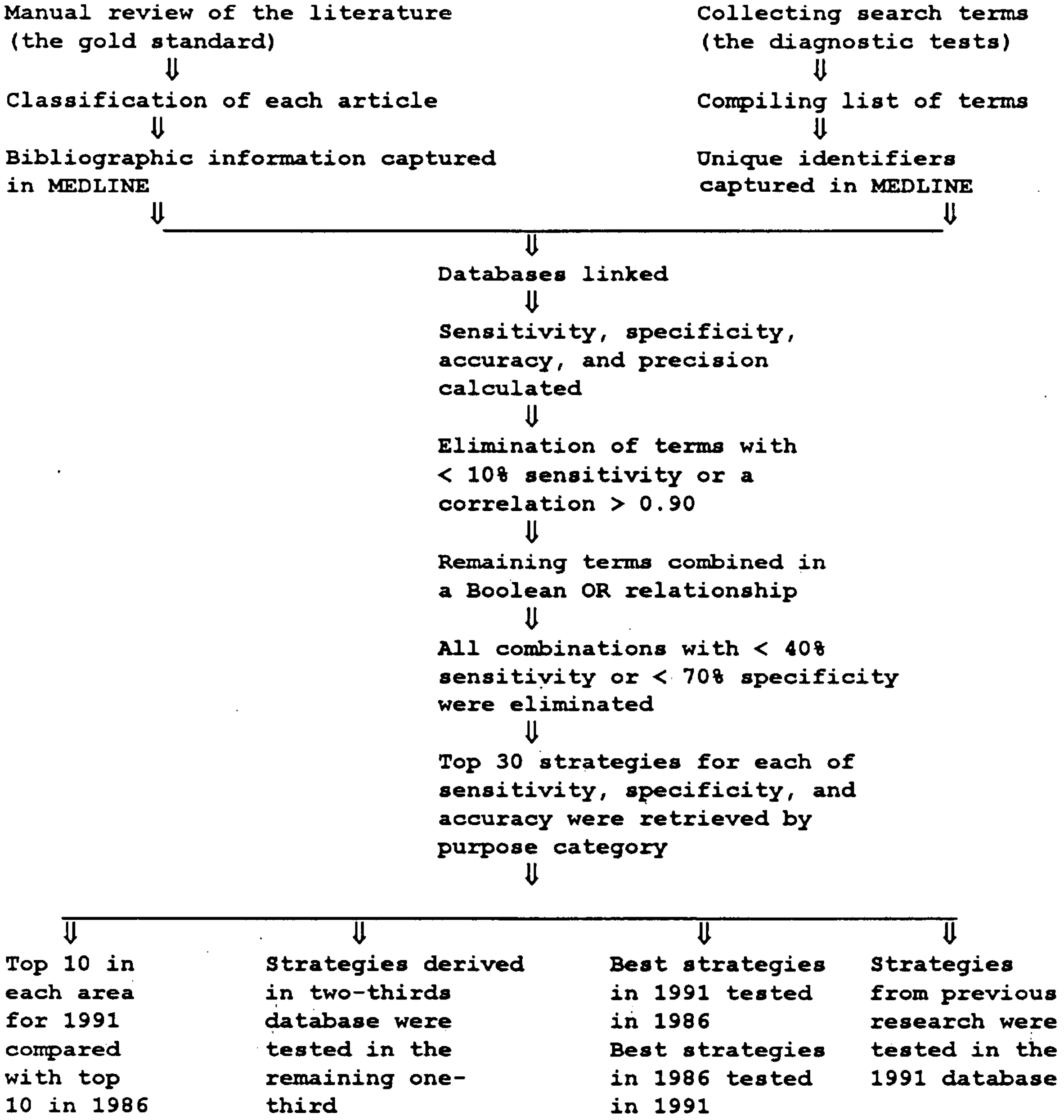
Manual review of the literature (the gold standard)
Classification of each article
Bibliographic information captured in MEDLINE
Collecting search terms (the diagnostic tests)
Compiling list of terms
Unique identifiers captured in MEDLINE
Databases linked
Sensitivity, specificity, accuracy, and precision calculated
Elimination of terms with < 10% sensitivity or a correlation > 0.90
Remaining terms combined in a Boolean OR relationship
All combinations with < 40% sensitivity or < 70% specificity were eliminated
Top 30 strategies for each of sensitivity, specificity, and accuracy were retrieved by purpose category
Top 10 in each area for 1991 compared with top 10 in 1986
Strategies derived in two-thirds database were tested in the remaining one-third
Best strategies in 1991 tested in 1986
Best strategies in 1986 tested in 1991
Strategies from previous research were tested in the 1991 database

Figure 1 Steps in data collection and analysis. In the left column, the steps involved in forming the database for the gold standard (manual review of the literature) are shown. In the right column, the steps involved in forming the database for the diagnostic tests (search strategies) are shown. In the center column, the databases are linked and data pertaining to the accuracy of the search strategies for identifying sound studies as defined by the "gold standard" are calculated. In the row across the bottom, how the data are reviewed and compared is shown.

servations, correspondence, news, obituaries, postgraduate and continuing-education forums, and notices.

To be considered of interest in the medical care of human adults, a 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 had to have been ≥18 years of age at study entry. Every format category was 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. Purpose categories and their corresponding definitions are shown in Table 3.

Studies in each purpose category were evaluated for methodologic rigor by determining whether they met one key methodologic criterion specific to their purpose as shown in Table 4. These criteria were based on critical appraisal criteria for applied research[4] but were set at a minimal level in recognition that few published studies meet the full set of criteria for unbiased clinical evaluation and that clinicians are likely to be better informed looking at the best available literature even if it falls short of perfection. For example, studies of prognosis were rated as meeting the methodologic criterion if they included a cohort of subjects who had the disease in question at baseline without the outcome of interest. This criterion would allow many more studies to qualify than would more stringent criteria such as that the studies include an inception cohort (or patients at a common and preferably early point in the courses of their disorders), with at least 80% follow-up and an objectively evaluated endpoint. Similarly, for assessment of diagnostic tests, the criterion was that data about the sensitivity and specificity of the test had to be reported; this is much less stringent than the standard of requiring independent or "blinded" interpretation of the test and the gold standard."

Table 2 ■

The Format Categories and Their Corresponding Definitions Used to Classify Journal Articles

Format	Definition
Original study	Any full-text article in which the investigators made firsthand observations
Review	Any full-text article that was bannered review, that had the word review in its title or in a section heading, or that indicated in the text that the intention was to review or summarize the literature about a topic
General article	A general or philosophical discussion of a topic without original observation and without 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
Conference report	Defined as such by the journal but reclassified by us as an original article or a review article when meeting those criteria
Decision analysis	Dissection of the management of patients into component parts, defining routes and consequences of management based on alternatives, for the purpose of defining optimal methods of management
Case report	An original study involving less than ten subjects

Table 3 ■

The Purpose Categories and Their Corresponding Definitions Used to Classify Journal Articles Classified as Original Studies, Reviews, or Case Reports* and of Interest

Purpose	Definition
Etiology	Content pertained directly to causation of a disease or condition
Prognosis	Content pertained directly to the prediction of the clinical course of the natural history of a disease with the disease existing at the beginning of the study
Diagnosis	Content pertained directly to the evaluation of a disease process, usually through comparing methods of arriving at a diagnosis
Treatment or prevention	Content pertained directly to therapy, prevention, or rehabilitation
Something else	Purpose of the study was something other than the above

*The terms original study, review, and case report are explained in Table 2.

Interrater reliability was assessed for the classification of articles for format, interest, purpose, and methods. In every case the degree of agreement beyond chance was assessed by the kappa statistic and was greater than 0.80 ($p < 0.05$).

The manual review of the literature served as the "gold standard" against which MEDLINE search strategies (the diagnostic tests) could be tested. Results of the study apply to original and review articles that are of acceptable quality from the perspective of applicability to clinical practice.

Collecting Search Terms

To construct a comprehensive set of search terms, we began a list of MeSH terms and textwords and then sought input from clinicians and librarians in the United States and Canada through interviews of known searchers; requests on electronic bulletin boards and in national publications, meetings, and conferences; and requests to the National Library of Medicine and the Canada Institute for Scientific and Technical Information. Individuals were asked what terms or phrases they used when searching for studies of etiology, prognosis, diagnosis, or therapy and for related review articles. Terms could be from MeSH, including publication types (pt), check tags, and subheadings (sh), or could be textwords (tw) denoting methodology in titles and abstracts of articles. The list, excluding incorrect MeSH terms, appears in the appendix. Some of the terms and phrases were different for the two years because new publication types were introduced in 1990 and 1991 and some of the corresponding terms changed definitions. Also, five terms retrieved no citations for the ten journals in 1986 and 1991, two terms retrieved no citations in 1991, and 21 terms retrieved no citations in 1986; these terms were discarded for the respective years.

Table 4 ■

Key Methodologic Criterion, According to Purpose Category,* Used to Determine the Methodologic Rigor of Journal Articles Classified for Purpose

Purpose	Key Methodologic Criterion
Etiology	Formal control group: random or quasi-random allocation of participants to treatment and control groups; or a nonrandomized concurrent control trial, a cohort analytic study with matching or statistical adjustment to create comparable groups, or a case–control study
Prognosis	A cohort of subjects who have the disease in question at baseline without the outcome of interest
Diagnosis	Provision of sufficient data to calculate the sensitivity and specificity of the test or likelihood ratios based on subjects who had been tested with both the test and the diagnostic standard
Treatment	Random or quasi-random allocation of participants to treatment and control groups
Review	Reproducible description of the methods for conducting the review (this criterion was applied to every review article regardless of the purpose for doing the review)

*Purpose categories and their definitions are provided in Table 3.

Table 5 ■

Classification of Original and Review Articles* That Appeared in Ten Journals† for 1986 and 1991 According to Four Purpose Categories and Whether They Met the Methodologic Criteria for These Categories, Based on Manual Review

Purpose Category‡	Year	
	1991	1986
Etiology	523§	531
Meeting methodologic criterion (%)	201 (38)	155 (29)
Prognosis	205	149
Meeting methodologic criterion (%)	133 (65)	106 (71)
Diagnosis	412	426
Meeting methodologic criterion (%)	111 (27)	92 (22)
Treatment	879	936
Meeting methodologic criterion (%)	281 (32)	270 (29)

*Original and review articles are defined in Table 2. There were 3,495 original and review articles for 1991 and 3,682 for 1986.
†The ten journals are listed in the Methods section.
‡Purpose categories are defined in Table 3. Their methodologic criteria are described in Table 4.
§Numbers other than those in parentheses are the numbers of articles.

Data Collection

Manual ratings of articles in the ten journals for 1986 and 1991 were recorded on data collection forms, and the bibliographic information, including eight-digit unique identifiers, for these articles was captured from MEDLINE. The manual review data for each article were double-entered into PARADOX by two independent data clerks. Each journal title was searched in MEDLINE for 1986 and 1991 and the publication types editorial, comment, letter, and news were excluded from the search using the Boolean AND NOT operator.

The MeSH terms and textwords to be tested were searched in MEDLINE for the ten journals for 1986 and 1991. The unique identifiers of retrieved citations were captured and then linked with the manual review data.

Testing Strategies

A computer program was written in Turbo Pascal to develop search strategies by creating all possible Boolean OR combinations of the terms in the appendix, and to determine the sensitivities, specificities, accuracy, and precision of the combinations of terms. To make the number of combinations more tractable, individual MeSH terms and textwords with a sensitivity <10% were eliminated, then terms with >0.90 correlation within each purpose category were determined and the term in the pair with the lower sensitivity was discarded from the combination analysis. Thus, the numbers of terms to combine for 1991 were 14 for etiology (16,383 combinations), 16 for prognosis (65,535 combinations), nine for diagnosis (511 combinations), and 15 for treatment (32,767 combinations). For 1986, the numbers of terms to combine were nine for etiology (511 combinations), 11 for prognosis (2,047 combinations), seven for diag-

Table 6 ■

Combinations of Terms (Medical Subject Headings and Textwords) with the Best Sensitivity for Detecting Sound Clinical Studies for Each Purpose Category

Purpose Category*	Search Strategy†	Sensitivity	Specificity	Precision
1991				
Etiology	Exp Cohort Studies or Exp Risk or Odds (tw) and Ratio: (tw) or Relative (tw) and Risk (tw) or Case (tw) and Control: (tw)	0.82 (0.65)‡	0.70	0.14
Prognosis	Incidence or Exp Mortality or Follow-up Studies or Mortality (sh) or Prognos: (tw) or Predict: (tw) or Course: (tw)	0.92 (0.75)§	0.73	0.11
Diagnosis	Exp Sensitivity a#d Specificity or Diagnosis& (px) or Diagnostic Use (sh) or Sensitivity (tw) or Specificity (tw)	0.92 (0.86)‡	0.73	0.09
Treatment	Randomized Controlled Trial (pt) or Drug Therapy (sh) or Therapeutic Use (sh) or Random: (tw)	0.99 (0.73)¶	0.74	0.22
1986				
Etiology	Exp Cohort Studies or Risk (tw) or Causation (tw) or Causal: (tw)	0.72 (0.78)‖	0.79	0.12
Prognosis	Prognosis or Exp Cohort Studies or Mortality (sh) or Natural (tw) and History (tw) or Predict: (tw) or Course (tw)	0.95 (0.92)‖	0.78	0.11
Diagnosis	Diagnosis& (px) or Specificity (tw)	0.86 (0.88)‖	0.73	0.07
Treatment	Random Allocation or Comparative Study or Drug Therapy (sh) or Placebo: (tw) or Controlled (tw) and Trial: (tw)	0.98 (0.80)‖	0.71	0.18

*Purpose categories are defined in Table 3.

†tw = textword; sh = subheading; px = subheading pre-explosion; pt = publication type.

‡Sensitivity in the 1986 database.

§This strategy could not be tested in the 1986 database. Strategy tested with 92% sensitivity in 1991 and 75% sensitivity in 1986: "Prognosis or Exp Morbidity or Follow-up Studies or Mortality (sh) or Prognos: (tw) or Clinical (tw) and Course (tw) or Predict: (tw) or Prognostic (tw) and Factor: (tw)."

¶This strategy could not be tested in the 1986 database. Strategy tested with 99% sensitivity in 1991 and 73% sensitivity in 1986: "Clinical Trial (pt) or Drug Therapy (sh) or Therapeutic Use (sh) or Random: (tw) or Double (tw) and Blind: (tw)."

‖Sensitivity in the 1991 database.

Table 7 ■

Combinations of Terms (Medical Subject Headings and Textwords) with the Best Specificity for Detecting Sound Clinical Studies for Each Purpose Category

Purpose Category*	Search Strategy†	Sensitivity	Specificity	Precision
1991				
Etiology	Cohort Studies or Case-control Studies	0.40	0.97 (0.98)‡	0.42
Prognosis	Prognosis or Survival Analysis	0.49	0.97 (0.97)§	0.34
Diagnosis	Exp Sensitivity a#d Specificity or Predictive (tw) and Value: (tw)	0.55	0.98 (0.99)¶	0.40
Treatment	Placebo: (tw) or Double (tw) and Blind: (tw)	0.57	0.97 (0.97)¶	0.56
1986				
Etiology	Exp Risk or Causation (tw) or Causal: (tw) or Relative (tw) and Risk (tw)	0.51	0.94 (0.88)‖	0.21
Prognosis	Prognosis or Natural (tw) and History (tw) or Prognostic (tw) and Factor: (tw)	0.58	0.97 (0.96)‖	0.31
Diagnosis	Specificity (tw) or Predictive (tw) and Value: (tw) or False (tw) and Positive (tw)	0.49	0.98 (0.96)‖	0.36
Treatment	Placebo: (tw) or Double (tw) and Blind: (tw)	0.58	0.97 (0.96)‖	0.60

*Purpose categories are defined in Table 3.

†tw = textword.

‡This strategy could not be tested in the 1986 database. Strategy tested with 94% specificity in 1991 and 98% specificity in 1986: "Odds (tw) and Ratio: (tw) or Relative (tw) and Risk (tw)."

§This strategy could not be tested in the 1986 database. Strategy tested with 97% specificity in the two years: "Prognosis or Prognostic (tw) and Factor:. (tw)."

¶Specificity in the 1986 database.

‖Specificity in the 1991 database.

nosis (127 combinations), and 14 for treatment (16,383 combinations).

Files containing the results for each individual combination (e.g., results for each of the 65,535 combinations of terms for prognosis in 1991) were sorted on each of sensitivity, specificity, and accuracy, in descending order. Using these sorted files, the 30 search strategies yielding the highest figures for each of sensitivity, specificity, and accuracy were reviewed for 1991 and for 1986 for all purpose categories. To compare search yields for 1986 and 1991, the ten strategies yielding the highest figures for each of sensitivity, specificity, and accuracy for each category for each year were compared and 95% confidence intervals (CIs) were calculated.

To determine the test–retest reliability of the search strategies, searches for treatment and etiology in 1991 and for treatment in 1986 were derived using a random two-thirds of the database and tested in the remaining one-third of the database. To assess the cross-year reliability of the searches, search strategies yielding the best values for each of sensitivity, specificity, and accuracy for each purpose category in 1991 were tested in the 1986 database and the best 1986 strategies were tested in the 1991 database. We also attempted to test search strategies described in previously published studies in our 1991 database.[8–14]

Results

The results of the manual review of the journals appear in Table 5. Even though the journals reviewed had the highest impact and immediacy ratings for general medicine and internal medicine and the criteria for methodologic adequacy were minimal, less than half of the articles met these criteria in three of the four categories. The exception, prognosis, may have had better results because our criterion was so lenient; very few of the studies would have met a stronger methodologic requirement of including an inception cohort.

Table 8 ■

Combinations of Terms (Medical Subject Heading and Textwords) with the Best Accuracy for Detecting Sound Clinical Studies for Each Purpose Category

Purpose Category*	Search Strategy†	Sensitivity	Specificity	Precision	Accuracy
1991					
Etiology	Cohort Studies or Exp Risk or Odds (tw) and Ratio: (tw) or Case (tw) and Control: (tw)	0.73	0.85	0.21	0.84 (0.51)‡
Prognosis	Survival Rate or Survival Analysis or Follow-up Studies or Prognos: (tw) or Clinical (tw) and Course (tw)	0.83	0.87	0.19	0.86 (0.87)§
Diagnosis	Exp Sensitivity a#d Specificity or Diagnosis (sh) or Diagnostic Use (sh) or Specificity (tw) or Predictive (tw) and Value: (tw)	0.86	0.84	0.13	0.84 (0.82)¶
Treatment	Randomized Controlled Trial (pt) or Random: (tw)	0.96	0.91	0.46	0.91 (0.91)‖
1986					
Etiology	Prospective Studies or Risk (tw)	0.70	0.84	0.14	0.83 (0.72)**
Prognosis	Prognosis or Follow-up Studies or Mortality (sh) or Natural (tw) and History (tw) or Prognostic (tw) and Factor: (tw) or Course (tw)	0.92	0.87	0.15	0.87 (0.81)**
Diagnosis	Diagnosis & (px) or Specificity (tw)	0.86	0.73	0.07	0.73 (0.76)**
Treatment	Random: (tw) or Placebo: (tw) or Double (tw) and Blind: (tw) or Controlled (tw) and Trial: (tw)	0.92	0.93	0.48	0.93 (0.88)**

*Purpose categories are defined in Table 3.

†tw = textword; sh = subheading; pt = publication type; px = subheading pre-explosion.

‡This strategy could not be tested in the 1986 database. Strategy tested with 83% accuracy in 1991 and 51% accuracy in 1986: "Exp Risk or Etiology& (px) or Cohort (tw) or Odds (tw) and Ratio: (tw)."

§This strategy could not be tested in the 1986 database. Strategy tested with 82% accuracy in 1991 and 87% accuracy in 1986: "Prognosis or Exp Mortality or Follow-up Studies or Mortality (sh) or Prognos: (tw) or Clinical (tw) and Course (tw) or Prognostic (tw) and Factor: (tw)."

¶Accuracy in the 1986 database.

‖This strategy could not be tested in the 1986 database. Strategy tested with 91% accuracy in the two years: "Random (tw) or Placebo: (tw)."

**Accuracy in the 1991 database.

For 1991 and 1986, the combinations of terms yielding the best sensitivities for the individual purpose categories are presented in Table 6. The sensitivities, specificities, and precision among the top 30 strategies for all purpose categories and the two years were similar to the figures presented here.

When comparing the top ten search strategies yielding the highest sensitivities in 1991 with the top ten search strategies for 1986, there was a 10% difference for etiology (CI, 1% to 20%), a 3% difference for prognosis (CI, −3% to 10%), a 6% difference for diagnosis (CI, −3% to 15%), and a 1% difference for treatment (CI, −1% to 4%).

For 1991 and 1986, the combinations of terms yielding the best specificities for the individual purpose categories are presented in Table 7. When reviewing the 30 search strategies yielding the highest specificities for the individual purpose categories and years, the drop in specificity across the top 30 ranged from 2% to 14%, with corresponding increases in sensitivity

Table 9 ■

Strategies Derived From Previous Research, Their Modifications for Testing in the 1991 Database, and Their Sensitivities

Author	Original Strategy	Modified Strategy	Sensitivity*
Poynard and Conn[9]	{Liver Disease or Biliary Tract Disease} and (Drug Therapy (sh) or Surgery (sh) or Radiotherapy (sh) or Therapy (sh)) and (Human and {Adult or Middle Age or Aged}) and (Comparative Study or Clinical Research)	Drug Therapy (sh) or Therapy (sh) or Comparative study	0.85
Dickersin et al.[10]	{Exp Hyperbilirubinemia and Exp Infant} or Random Allocation or Exp Research Design or Clinical Trials or Random: (tw)	Random Allocation or Exp Research Design or Clinical Trials or Random: (tw)	0.94
Kirpalani et al.[8]	Therapy (sh) or Prevention and Control (sh) or Exp Feeding Methods or Diet Therapy (sh) or Drug Therapy (sh) or Random Allocation or All Random: (tw) or Placebos	Therapy (sh) or Prevention and Control (sh) or Drug Therapy (sh) or Random Allocation or Random: (tw) or Placebos	0.98
Schuyler et al.[11]	Clinical Trials or Clinical Trial (pt)	Clinical Trials or Clinical Trial (pt)	0.93
Gotzsche and Lange[12]	{Exp Arthritis, Rheumatoid and ≥1 of 17 NSAIDS} or Comparative Study or Double-blind Method or Double and Blind (tw)	Comparative Study or Double-blind Method or Double (tw) and Blind: (tw)	0.73
Bernstein[14]	{Liver or Liver Disease or Biliary Tract or Biliary Tract Diseases} and (Human and {Adult or Middle Age or Aged}) and (Random Allocation or Double-blind Method) or (Clinical Trials or Comparative Study and {Random (tw) or Double Blind (tw) or Controlled (tw)})	Random Allocation or Double-blind Method or Clinical Trials or Comparative Study or Random: (tw) or Double (tw) and Blind: (tw)	0.96
Jadad and McQuay[13]	({Pain* or Exp Pain or Analg* or Exp Analgesia or Exp Analgesics} and Human) and ({Exp Clinical Trials or Clinical Trials or Random* or Random Allocation or Randomized Controlled Trials or Double Blind or Double-blind Method or Meta-analysis} and Human)	Exp Clinical Trials or Clinical Trials or Random: (tw) or Random Allocation or Randomized Controlled Trials or Double (tw) and Blind: (tw) or Double-blind Method	0.94

*sh = subheading; tw = textword; pt = publication type; NSAIDS = nonsteroidal anti-inflammatory drug. By comparison, our most sensitive strategy for therapy (from Table 6) had a sensitivity of 99%.

from 9% to 35%. For example, for treatment in 1991, specificity dropped from 97% to 94% across the top 30 but sensitivity increased from 57% to 92%.

When comparing the top ten search strategies yielding the highest specificities in 1991 with the top ten strategies in 1986, differences ranged from 1% (CI, 0.1% to 3%) to 5% (CI, 4% to 7%) for etiology. All comparisons for prognosis and diagnosis showed a 1% difference, and the comparison for treatment revealed that in most cases the specificities were the same.

For 1991 and 1986, the combinations of terms yielding the best accuracies for the individual purpose categories are presented in Table 8. The accuracies were the same among the top 30 search strategies for all purpose categories and the two years.

Differences in accuracy for 1991 compared with 1986 favored 1991 for etiology and diagnosis. The differences were small for all categories except diagnosis. When comparing the top ten search strategies yielding the highest accuracies in 1991 with the top ten strategies in 1986, the differences for etiology ranged from 0.8% (CI, −1% to 3%) to 1% (CI, −0.5% to 3%). For prognosis, the differences ranged from 1% (CI, −0.5% to 3%) to 9% (CI, 7% to 11%). For diagnosis, all differences were 11% (CI, 9% to 12%). For treatment, all differences were 2% (CI, 0.4% to 3%).

The top ten search strategies for each of sensitivity, specificity, and accuracy for treatment in 1991 and 1986 and for etiology in 1991 were derived in two-thirds of the database and tested in the remaining one-third of the database. The largest difference was for the comparison of the ten strategies yielding the highest values for sensitivity for etiology in 1991. There was a 9% difference for all ten strategies, but this was not statistically significant (82% for the derived database vs. 73% for the test database; CI for the difference, −4% to 22%). In most cases the values were identical or within 1% to 2%, and none was statistically significant.

The strategies yielding the best sensitivities, specificities, and accuracy in 1991 were tested in the 1986 database for all purpose categories, given that the terms were available in that year (Tables 6, 7, and 8). The best 1986 strategies were also tested in the 1991 database. Seven (58%) of the 12 best strategies of 1991 could not be run in the 1986 database because terms used in 1991 were not available in 1986. When the best search strategies for 1991 that could be run in the 1986 database were tested, sensitivity and accuracy usually decreased substantively and specificity remained about the same or increased slightly. For example, the strategy yielding the best sensitivity in 1991 for etiology (82%) had a 65% sensitivity in 1986 (CI for the 17% difference, 7% to 26%; Table 6). When testing the best 1986 search strategies in the 1991 database, in most cases sensitivity, specificity, and accuracy decreased slightly.

Search strategies derived in previously published studies were tested in our 1991 database. All searches were for studies of treatment. Not all terms could be included because our database did not contain disease and content terms or some of the method terms. The strategies derived in previous research and variations of these strategies that could be tested in our database are shown in Table 9, along with their corresponding sensitivities.

Discussion

Our findings show that some search strategies can achieve high sensitivity and specificity for detecting sound clinical studies in MEDLINE. The sensitivity of methodologic search terms in MEDLINE was enhanced by combining MeSH terms and textwords when attempting to retrieve studies meeting methodologic criteria for the etiology, prognosis, diagnosis, prevention, and treatment of disorders in adult general medicine. The gains for prognosis for 1991 and 1986 were substantial, with lesser but still statistically significant gains for diagnosis and etiology. Near-perfect sensitivities of 99% and 98% were achieved for treatment for 1991 and 1986, respectively. As would be expected, however, there was an inverse relationship between sensitivity and specificity, leading to a decrease in precision. When sensitivity was maximized, precision fell to below 25% in all cases.

The specificities of several individual methodologic MeSH terms and textwords were near perfect, but this was at the expense of sensitivity. When combining terms, near-perfect specificity was maintained in numerous strategies with modest increases in sensitivity. For treatment, individual terms outperformed the combinations for maximizing specificity while maintaining sensitivity for both 1991 and 1986.[6]

Search strategies were also derived that maximized accuracy. Substantive increases in accuracy were also achieved by combining terms.

In all cases, search strategies for treatment outperformed strategies derived for the other purpose categories. The search strategy that yielded the best sensitivity (99%) for treatment in 1991 was Randomized Controlled Trial (pt) or Drug Therapy (sh) or

Therapeutic Use (sh) or All Random: (tw). This sensitivity is much higher than those reported in previous studies.[8-14] However, a direct comparison between search strategies developed in previous studies and the strategies developed in our study is speculative for a number of reasons, including the lack of content terms in our database, differences in methods terms, and differences in the years and journals accessed. With these limitations in mind, some further observations appear warranted.

When modified to contain only methods terms that appeared in our database, six of the seven search strategies derived in previously published studies yielded higher sensitivities in our 1991 database than they did in the original studies. This difference may be due to better indexing for priority journals, an improvement in indexing over the years, or higher sensitivity for search strategies that do not contain disease and content terms. For example, the search strategy derived by Kirpalani et al.[8] (Table 9), when modified, did not contain the terms Exp Feeding Methods and Diet Therapy (sh) because these content terms were not available in our database. One strategy, developed by Gotzsche and Lange,[12] had a lower sensitivity in our database. When searching for double-blind trials of nonsteroidal anti-inflammatory drugs (NSAIDS) in rheumatoid arthritis published before 1985, they found a 98% sensitivity for {Exp Arthritis, Rheumatoid and ≥1 of 17 NSAIDS} or Comparative Study or Double-blind Method or Double (tw) and Blind (tw). When content terms were excluded, the sensitivity for this strategy was 73% in our 1991 database. The difference in sensitivity may be due to the types of articles that were the targets of the searches: Gotzsche and Lange were searching for double-blind trials, whereas we were searching for randomized controlled trials.

We had a prescreening step in the development of our search strategies. When searching for each journal title in MEDLINE, the publication types editorial, comment, letter, and news were excluded from the search using the Boolean AND NOT operator. This prescreening step would have no effect on the sensitivities calculated for the combinations of terms because studies meeting methodologic criteria were defined by the manual review of the literature. This step would, however, result in the overestimation of specificity and precision. Thus, searchers would be advised to include this prescreening step if maintaining similar levels of specificity and precision are of concern.

The search strategies presented here can aid searchers, particularly clinicians who are inexperienced in constructing complex searches, in retrieving studies that meet at least one major criterion for scientific merit for applied health care research while filtering out studies with weaker designs. Such strategies are bound to retrieve some false-positive articles and miss others that should be retrieved. One major methodologic criterion was chosen for each purpose category in order to keep the filters simple. As such, retrieved articles would have to be further evaluated by the user to determine their methodologic soundness and clinical applicability. Unless even more elaborate strategies are developed, false-negative articles (i.e., appropriate articles not retrieved by a search) can be retrieved only by hand searching journals or through other labor-intensive means.

One limitation of our study was that only priority journals were included in the search. One strength of the study was the highly reproducible classification of articles in the manual review of the literature, which served as the "gold standard." Another strength was the validation of the search strategies for etiology for 1991 and for treatment for 1991 and 1986.

Search strategies derived to maximize sensitivity and accuracy for treatment outperformed strategies derived for the other purpose categories. Search strategies derived to maximize specificity were comparable across purpose categories. Publication types are clearly valuable when searching for treatment studies and they are needed when searching for etiology, prognosis, and diagnosis studies.

Because the testing of the 1991 search strategies in the 1986 database revealed a statistically significant difference in most cases, back file searches should be modified appropriately. For example, when searching for studies of treatment in 1991 with the intention of maximizing sensitivity, the methodologic filter Randomized Controlled Trial (pt) or Drug Therapy (sh) or Therapeutic Use (sh) or All Random: (tw) should be included. However, when searching in 1986 this filter should be changed to Random Allocation or Comparative Study or Drug Therapy (sh) or All Placebo: (tw) or Controlled (tw) and Trial (tw).

Further research is needed to address how these methodologic search filters perform when disease and content terms are added and when they are applied using nonpriority journals; to develop optimal search strategies that could be made available to clinical end users of MEDLINE; to develop search strategies that maximize sensitivity while preserving precision; to test these strategies in a real searching environment; and to determine whether there is any interaction between content and methods terms.

References ■

1. Dr. Lindberg reports to the Congress: 1991—the year of outreach. Gratefully Yours. March/April 1992:4–5.
2. Lancaster FW. MEDLARS: report on the evaluation of its operating efficiency. Am Documentation. 1969;119–42.
3. Haynes RB, Johnston ME, McKibbon KA, Walker CJ, Willan AR. A randomized controlled trial of a program to enhance clinical use of MEDLINE. Online J Curr Clin Trials. 1993 (Doc No 56).
4. Sackett DL, Haynes RB, Guyatt GH, Tugwell P. Clinical Epidemiology: A Basic Science for Clinical Medicine, 2nd ed. Boston: Little, Brown, 1991.
5. 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.
6. Wilczynski NL, Walker CJ, McKibbon KA, Haynes RB. Assessment of methodologic search filters in MEDLINE. Proc Annu Symp Comp Appl Med Care. 1994;17:601–5.
7. Science Citation Index, vol. 16. Journal Citation Reports, 1984. Philadelphia: Institute for Scientific Information, 1985.
8. 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.
9. Poynard T, Conn HO. The retrieval of randomized clinical trials in liver disease from the medical literature: a comparison of MEDLARS use and manual methods. Control Clin Trials. 1985;6:271–9.
10. Dickersin K, Hewitt P, Mutch L, Chalmers I, Chalmers TC. Perusing the literature: comparison of MEDLINE searching with a perinatal trials database. Control Clin Trials. 1985;6:306–17.
11. Schuyler P, Dickersin K, Scherer R, Wright N. Identification of randomized clinical trials using MEDLINE (abstract). Second Meeting of the International Congress on Peer Review, September 9–11, 1993, Chicago, Illinois.
12. Gotzsche PC, Lange B. Comparison of search strategies for recalling double-blind trials from MEDLINE (abstract). Dan Med Bull. 1991;38:476–8.
13. 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).
14. Bernstein F. The retrieval of randomized clinical trials in liver diseases from the medical literature: manual versus MEDLARS searches. Control Clin Trials. 1988;9:23–31.

APPENDIX

Complete List of Search Terms

***Notes*: Terms that retrieved no citations in 1991 are marked with *; terms that retrieved no citations in 1986 are marked with †; terms with <10% sensitivity in 1991 are marked with ‡; terms with <10% sensitivity in 1986 are marked with §; terms with >0.90 correlation with another term in 1991 are marked with ¶; terms with >0.90 correlation with another term in 1986 are marked with **; subheading (sh) pre-explosion is noted by & (px); and truncation is noted by :.**

Etiology

MeSH terms: exp case control studies§; case control studies†; retrospective studies‡§; exp cohort studies; cohort studies†; exp longitudinal studies ¶**; longitudinal studies‡§; follow-up studies§; prospective studies; cross-sectional studies‡§; exp causality†¶; causality†‡; risk factors †¶; exp risk; risk‡**; logistic models†‡; odds ratio†‡; etiology&; etiology (sh).

Textwords: cohort§; risk; all etiol: or all aetiol:‡§; odds and all ratio:§; causation or all causal:‡; relative and risk; case and all control:; case and comparison‡§; case and referent*†.

Prognosis

MeSH terms: exp cohort studies; cohort studies†‡; exp longitudinal studies ¶**; longitudinal studies‡§; follow-up studies; prospective studies; prognosis; exp morbidity§; morbidity‡§; incidence†; exp mortality§; mortality‡§; cause of death†‡; infant mortality‡§; maternal mortality‡§; survival rate†; survival analysis†; mortality (sh).

Textwords: natural and history‡; all prognos:; inception and cohort†‡; clinical and course§; all predict:; all outcome:; clinical and all consequence:‡§; prognostic and all factor:; morbidity‡§; course.

Diagnosis

MeSH terms: exp sensitivity and specificity§; sensitivity and specificity§; predictive value of tests§; ROC curve†‡; exp diagnostic errors‡§; diagnostic errors‡§; false positive reactions‡§; false negative reactions‡§; diagnosis, differential‡§; diagnosis&; diagnosis (sh); diagnostic use (sh).

Textwords: sensitivity; specificity; predictive and all value:; post and test and all probabilit:*§; post and test and likelihood*§; likelihood and all ratio:‡§; false and rate‡§; false and positive‡; false and negative‡§; receiver and all operat: and characteristic§; ROC‡§; independent and comparison‡§; all mask: and comparison†‡; all blind: and comparison‡§; gold and standard‡§; pre and test and all probabilit:*†; pre and test and likelihood*; independent comparison*†.

Treatment

MeSH terms: exp research design; research design‡§; double-blind method¶**; random allocation‡; exp clinical trials‡; clinical trials‡**; exp multicenter studies*†; multicenter studies†‡; randomized controlled trials†‡; clinical trial (pt); multicenter study (pt)†; randomized controlled trial (pt)†; comparative study; single-blind method†‡; placebos‡§; prevention & control (sh); therapy&; therapy (sh); drug therapy (sh); therapeutic use&; therapeutic use (sh).

Textwords: all random:; all placebo:; double and all blind:; all mask:‡§; single and all blind:‡§; controlled and all trial:.