Robustness of Empirical Search Strategies for Clinical Content in MEDLINE

Nancy L. Wilczynski, MSc, R. Brian Haynes, MD, PhD, for the Hedges Team
Health Information Research Unit, McMaster University, Hamilton, Ontario, Canada

Abstract

Background: 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 to improve the accuracy of searching for such studies. Methodologic search filters have been tested in previous MEDLINE files but indexing continues to evolve and the operating characteristics of these search filters in current MEDLINE files are unknown.
Objective: To determine the robustness of empirical search strategies developed in 1991 for detecting clinical content in MEDLINE in the year 2000.
Design: A survey based on a hand search of 171 core health care journals using predetermined quality indicators for scientific merit and clinical relevance.
Methods: 6 trained, experienced research assistants read all issues of 171 journals for the publishing year 2000. Each article was rated using purpose and quality indicators and categorized into clinically relevant original studies, review articles, general papers, or case reports. The original and review articles were then categorized as 'pass' or 'fail' for methodologic rigor in the areas of therapy/quality improvement, diagnosis, prognosis, causation, economics, clinical prediction, and qualitative and review articles. Search strategies developed in 1991 were tested in the 2000 database.
Main outcome measures: Sensitivity, specificity, precision, and accuracy of the search strategies.
Results: Search strategies developed in 1991 generally performed at least as well in 2000 for both best single terms and combinations of terms for high-sensitivity MEDLINE searches for studies of treatment, prognosis, etiology and diagnosis. For example, the accuracy of "clinical trial (pt)" rose from 91.6% to 94.4% ($P<0.05$) for retrieving high-quality studies of treatments.
Conclusion: Most MEDLINE search strategies developed in 1991 are robust when searching in the publishing year 2000.

Introduction

Health care research dissemination suffers from both "supply" and "demand" problems. On the supply side, advances in health care practice are published in a wide array of journals. Journals are searchable through electronic databases (eg, MEDLINE) but retrieval problems for clinical end-users are multiplied by the very low concentration of studies that are new, sound, and ready for application [1]. On the demand side, practitioners have difficulty keeping up with new advances in health care [2,3], most researchable information needs are unmet [4], and practitioner's searches lack sensitivity, specificity, and precision [5]. If large electronic databases are to be helpful to clinical end-users, end-users must be able to retrieve articles that are scientifically sound and directly relevant to the health problem they are trying to solve, without missing key studies or retrieving excessive numbers of irrelevant or misleading studies. The use of methodologic search filters has been advocated, [6] and filters have been developed, to improve the accuracy of searching for such studies [7,8].

One possible solution to these problems is to develop "methodologic search filters" to improve the retrieval of clinically relevant and scientifically sound studies from large, general purpose, biomedical research bibliographic databases, such as, MEDLINE. For example, in MEDLINE, filters are created by adding, to the usual disease content terms, Medical Subject Headings (MeSH), explosions (px), publication types (pt), subheadings (sh) and textwords (tw) that detect research design features indicating methodologic rigor for applied health care research, for instance, 'Exp myocardial infarction and (randomized controlled trial (pt) or clinical trial (pt))'. In the early 1990s, our group developed search filters for MEDLINE on a small subset of journals and for 4 types of journal articles [9,10], and these strategies have been adapted for use in the Clinical Queries interface of MEDLINE (http://www.ncbi.nlm.nih.gov:80/entrez/query/static/clinical.html). This research is being updated and expanded using data from the publishing year 2000.

When developing search strategies in the early 1990s we found that strategies that maximized sensitivity or specificity in 1991 had to be modified when back searching in 1986. Partially, the modification was

necessary because methodologic publication types were introduced in 1990 (eg, clinical trial (pt)). Since modifications were required when back searching we questioned the performance of these search strategies when searching forward in time. The purpose of this report is to show how well search strategies developed in MEDLINE in 1991 perform in the publishing year 2000.

Methods

To determine the information retrieval properties of search strategies developed in 1991 they were treated as "diagnostic tests" or screening procedures for the detection of relevant citations in 2000. Borrowing from the concepts of diagnostic test evaluation and library science, the sensitivity, specificity, precision, and accuracy of the MEDLINE searches were calculated. 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. Sensitivity and specificity are not affected by the proportion of high quality articles in the database; precision is dependent on this proportion, and so is accuracy, but to a lesser extent. The yield of 1991 MEDLINE search filters were determined by comparison with manual hand searches of journals in 2000, the gold standard.

For the year 2000 six research associates (compared with three in 1991) reviewed 171 journal titles (compared with 10 peer reviewed general adult medicine journals in 1991) and applied methodologic criteria (Table 1) to each item in each issue to determine if the article was methodologically sound for the purpose categories (8 in 2000 compared with 4 in 1991) listed in Table 1. The 171 journal titles reviewed in 2000 were chosen in an iterative process based on recommendations of clinicians and librarians, Science Citation Index Impact Factors, 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.
The methodologic criteria applied in 2000 were more rigorous than in 1991 and are outlined in Table 1 for the categories that were the same for these 2 studies. Research staff were rigorously calibrated and inter-rater agreement for application of methodologic criteria exceeded 80% beyond chance for all study purpose categories [11].

Results

There were 25,001 articles categorized as original studies or review articles in the 171 journal titles for the year 2000. Table 2 shows the best single terms for high-sensitivity MEDLINE searches that were developed in 1991 and the operating characteristics of these terms in 1991 and 2000, both in the 171 journal set, and in the same 10 journal subset reviewed in 1991. Best single terms derived in 1991 performed at least as well, and usually significantly better for sensitivity, specificity and accuracy in both the 171 journal set and the original 10 journal subset. For example, the accuracy of "clinical trial (pt)" for studies of treatments improved from 91.6% in 1991 to 94.4% for the full journal set in 2000 and 96.2% in the 10 journal subset, a rise of 2.8% (95% confidence interval (CI) 1.9% to 3.8%) and 4.6% (CI 3.6% to 5.5%), respectively.

In contrast, with only one exception, precision (the proportion of retrieved articles that are of high quality) was less in the year 2000 than in 1991 (Table 2) (see Discussion).

Table 3 shows the operating characteristics for the combination of terms with the best sensitivity for 1991. As expected, combinations increased sensitivity. In almost all cases sensitivity was better in 2000 than 1991 in both the 171 and 10 journal sets, although not significantly so; for specificity and accuracy, the results were significantly better in 2000 than 1991 for both journal sets, but precision was significantly lower.

Results were less consistent using the combination of terms with the best specificity derived in 1991 (Table 4). The results were trivially different for specificity and accuracy. Sensitivity was significantly less for treatment and etiology, but significantly better for diagnosis in the full 2000 database. Precision was lower once again in 2000 except for studies of treatment in the 10 journal set.

Discussion

Of most interest to clinical searchers is the robustness of the best single and combined terms for high-sensitivity MEDLINE searches for studies of treatment, prognosis, etiology and diagnosis. These terms performed at least as well for sensitivity, specificity, and accuracy in 1991 and 2000. The improvements in sensitivity, specificity and accuracy

observed in 2000 searches are likely due to increasing the rigor of the methodologic criteria applied in 2000, compared with 1991. Presumably, higher quality studies are better reported and more accurately indexed.

The precision of searching was generally less, and often substantially so, in 2000. This is also mainly an expected effect of increasing the rigor of the standards used to define articles of high quality in the 2000 review. Precision is affected by the proportion of high quality articles in the database (whereas sensitivity and specificity are not), and higher standards reduced the proportion of articles rated as meeting the standard. Therapy articles were relatively spared in the change in precision as the

Table 1 – Purpose categories and methodologic rigor for the hand search of the literature

Purpose	Methodologic Rigor
Therapy/ Quality Improvement	**2000** - Random allocation of participants to comparison groups; Outcome assessment of at least 80% of those entering the investigation; Analysis consistent with study design. **1991** - Random or quasi-random allocation of participants to comparison groups.
Diagnosis	**2000** - Inclusion of a spectrum of participants; Objective diagnostic ("gold") standard OR current clinical standard for diagnosis; Participants received both the new test and some form of the diagnostic standard; Interpretation of diagnostic standard without knowledge of test result and visa versa; Analysis consistent with study design. **1991** - 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 gold standard.
Prognosis	**2000** - 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; Analysis consistent with study design. **1991** - A cohort of individuals who have the disease at baseline without the outcome of interest.
Causation (Etiology)	**2000** - Observations concerned with the relationship between exposures and putative clinical outcomes; Data collection is prospective; Clearly identified comparison group(s); Blinding of observers of outcome to exposure. **1991** - Clearly identified comparison group(s).

Table 2 – Best single terms for high-sensitivity MEDLINE searches derived in 1991 compared with performance in 2000

Search strategy	Sensitivity (%) 1991 / 2000[a] / 2000[b] Difference (95% CI)† Difference (95% CI)‡	Specificity (%) 1991 / 2000[a] / 2000[b] Difference (95% CI)† Difference (95% CI)‡	Precision (%) 1991 / 2000[a] / 2000[b] Difference (95% CI)† Difference (95% CI)‡	Accuracy (%) 1991 / 2000[a] / 2000[b] Difference (95% CI)† Difference (95% CI)‡
For studies of treatment Clinical trial (pt)	92.5 / 94.8/ 97.9 2.3 (-0.8 to 6.1) 5.4 (2.2 to 9.2)*	91.6 / 94.4 / 96.1 2.8 (1.9 to 3.9)* 4.5 (3.5 to 5.6)*	48.6 / 35.8 / 50.1 -12.8 (-17.5 to -8.2)* 1.5 (-3.9 to 6.9)	91.6 / 94.4 / 96.2 2.8 (1.9 to 3.8)* 4.6 (3.6 to 5.5)*
For studies of prognosis Exp cohort studies	60.2 / 50.0 / 63.6 -10.2 (-22.4 to 2.4) 3.4 (-18.9 to 22.6)	80.8 / 86.3 /92.0 5.5 (4.1 to 6.8)* 11.2 (9.8 to 12.6)*	10.9 / 1.4 / 1.6 -9.5 (-12.1 to –7.5)* -9.3 (-11.9 to –7.1)*	80.0 / 86.1 / 92.0 6.1 (4.7 to 7.5)* 12.0 (10.5 to 13.4)*
For studies of etiology Risk (tw)	67.2 / 71.8 / 76.1 4.6 (-4.7 to13.9) 8.9 (-2.7 to 19.4)	79.4 / 87.5 / 88.5 8.1 (6.6 to 9.5)* 9.1 (7.6 to 10.6)*	14.8 / 3.3 /5.1 -11.5 (-14.0 to –9.2)* -9.7 (-12.4 to –7.2)*	78.7 / 87.4 / 88.4 8.7 (7.4 to 10.2)* 9.7 (8.2 to 11.2)*
For studies of diagnosis Sensitivity (tw)	56.8 / 70.1 / 57.1 13.3 (1.5 to 25.0)* 0.3 (-26.1 to 24.3)	96.6 / 97.4 / 98.3 0.8 (0.2 to 1.4)* 1.7 (1.0 to 2.4)*	32.8 / 7.4 / 4.1 -25.4 (-32.5 to –19.0)* -28.7 (-36.1 to –21.6)*	95.4 / 97.3 / 98.2 1.9 (1.2 to 2.7)* 2.8 (2.2 to 3.6)*

*Difference is statistically significant. 2000[a] = search tested in 2000 database on all 171 journals reviewed. 2000[b] = search tested in 2000 database on only the 10 journal subset reviewed in 1991. †Comparing 1991 and 2000[a]. ‡Comparing 1991 and 2000[b].

Table 3 – Combination of terms with the best sensitivity in MEDLINE; searches derived in 1991 and compared with performance in 2000

Search strategy	Sensitivity (%) 1991 / 2000[a] / 2000[b] Difference (CI)† Difference (CI)‡	Specificity (%) 1991 / 2000[a] / 2000[b] Difference (CI)† Difference (CI)‡	Precision (%) 1991 / 2000[a] / 2000[b] Difference (CI)† Difference (CI)‡	Accuracy (%) 1991 / 2000[a] / 2000[b] Difference (CI)† Difference (CI)‡
For studies of treatment randomized controlled trial (pt) OR drug therapy (sh) OR therapeutic use (sh) OR random: (tw)	98.9 / 98.3 / 99.8 -0.6 (-2.0 to 2.3) 1.0 (-0.4 to 3.7)	74.2 / 79.2 / 77.9 5.0 (3.2 to 7.0)* 3.7 (1.8 to 5.8)*	22.0 / 13.4 / 15.3 -8.6 (-11.6 to -5.7)* -6.7 (-10.0 to –3.7)*	76.1 / 79.8 / 78.8 3.7 (2.0 to 5.5)* 2.7 (0.8 to 4.6)*
For studies of prognosis incidence OR exp mortality OR follow-up studies OR mortality (sh) OR prognos: (tw) OR predict: (tw) OR course: (tw)	91.7 / 91.9 / 100.0 0.2 (-7.3 to 7.3) 8.3 (-6.7 to 14.2)	72.7 / 78.1 / 83.7 5.4 (3.8 to 7.0)* 11.0 (9.3 to 12.6)*	11.0 / 1.6 / 1.2 -9.4 (-11.4 to –7.7)* -9.8 (-11.8 to –7.9)*	73.5 / 78.2 / 83.7 4.7 (3.2 to 6.2)* 10.2 (8.6 to 11.9)*
For studies of 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)	81.7 / 84.5 / 86.4 2.8 (-5.5 to 11.4) 4.7 (-5.6 to 14.0)	70.2 / 76.8 / 81.7 6.6 (4.7 to 8.7)* 11.5 (9.5 to 13.6)*	14.0 / 2.1 / 3.7 -11.9 (-14.4 to –9.6)* -10.3 (-12.9 to –7.9)*	70.9 / 76.9 / 81.7 6.0 (4.2 to 8.0)* 10.8 (8.9 to 12.9)*
For studies of diagnosis exp sensitivity a#d specificity OR sensitivity (tw) OR diagnosis& (sh) OR diagnostic use (sh) OR specificity (tw)	91.9 / 96.6 / 92.9 4.7 (-0.9 to 11.7) 1.0 (-23.8 to 10.3)	72.9 / 65.6 / 78.2 -7.3 (-8.9 to –5.8)* 5.3 (3.5 to 6.9)*	9.0 / 0.8 / 0.5 -8.2 (-9.9 to –6.6)* -8.5 (-10.3 to –6.9)*	73.6 / 65.7 / 78.2 -7.9 (-9.4 to –6.4)* 4.6 (3.0 to 6.3)*

*Difference is statistically significant. 2000[a] = search tested in 2000 database on all 171 journals reviewed. 2000[b] = search tested in 2000 database on only the 10 journal subset reviewed in 1991. †Comparing 1991 and 2000[a], 95% CI. ‡Comparing 1991 and 2000[b], 95% CI.

Table 4 – Combination of terms with the best specificity in MEDLINE; searches derived in 1991 and compared with performance in 2000

Search strategy	Sensitivity (%) 1991 / 2000[a] / 2000[b] Difference (CI)† Difference (CI)‡	Specificity (%) 1991 / 2000[a] / 2000[b] Difference (CI)† Difference (CI)‡	Precision (%) 1991 / 2000[a] / 2000[b] Difference (CI)† Difference (CI)‡	Accuracy (%) 1991 / 2000[a] / 2000[b] Difference (CI)† Difference (CI)‡
For studies of treatment double (tw) and blind: (tw) OR placebo: (tw)	56.9 / 42.3 / 53.1 -14.6 (-22.3 to –6.7)* -3.8 (-12.3 to 4.9)	96.5 / 98.1 / 98.7 1.6 (0.8 to 2.4)* 2.2 (1.5 to 3.1)*	55.9 / 42.2 / 62.5 -13.7 (-21.5 to –5.9)* 6.6 (-2.1 to 15.3)	93.5 / 96.3 / 96.9 2.8 (1.9 to 3.9)* 3.4 (2.5 to 4.6)*
For studies of prognosis prognosis OR survival analysis	48.9 / 39.3 / 36.4 -9.6 (-21.8 to 2.9) -12.5 (-31.6 to 10.0)	96.5 / 95.1 / 96.2 -1.4 (-2.1 to –0.1)* -0.3 (-1.1 to 0.4)	34.0 / 3.0 / 1.9 -31.0 (-38.0 to –24.6)* -32.1 (-39.2 to –25.6)*	94.8 / 94.9 / 96.1 0.1 (-1.0 to 0.6) 1.3 (0.05 to 2.2)*
For studies of etiology case-control studies OR cohort studies	40.1 / 25.9 / 26.1 -14.2 (-24.6 to –3.9)* -14.0 (-25.6 to –1.3)*	96.5 / 95.5 / 96.3 -1.0 (-1.8 to –0.2)* -0.2 (-1.0 to 0.7)	41.9 / 3.3 / 5.4 -38.6 (-47.1 to –30.6)* -36.5 (-45.1 to –28.2)*	93.1 / 95.1 / 95.7 2.0 (1.0 to 3.1)* 2.6 (1.6 to 3.8)*
For studies of diagnosis exp sensitivity a#d specificity OR predictive (tw) and value: (tw)	54.8 / 79.6 / 64.3 24.8 (13.2 to 35.8)* 9.5 (-18.0 to 31.4)	98.0 / 94.9 / 96.6 -3.1 (-3.6 to –2.7)* -1.4 (-2.0 to –0.8)*	39.9 / 4.5 / 2.4 -35.4 (-43.4 to –27.9)* -37.5 (-45.6 to –29.9)*	96.6 / 94.8 / 96.5 -1.8 (-2.4 to –1.1)* -0.1 (-0.7 to 0.6)

*Difference is statistically significant. 2000[a] = search tested in 2000 database on all 171 journals reviewed. 2000[b] = search tested in 2000 database on only the 10 journal subset reviewed in 1991. †Comparing 1991 and 2000[a], 95% CI. ‡Comparing 1991 and 2000[b], 95% CI.

quality standard did not change much (Table 1).

Changes in precision could also be explained by a larger number of lower quality journals in the 171 dataset [12], but this does not appear to be the case, as the differences in precision between the 171 and 10 journal sets in 2000 are not consistent.

While low precision in searching can be of concern, the low values here should not be over-interpreted: we did not limit the searches by clinical content terms, as would be the usual case in clinical searches. We have also not tested "and" and "and not" combinations, which would be expected to increase the specificity of searching, a major determinant of precision. Limiting by clinical journal subsets may also provide protection against low precision [12]. The next phases of our project will focus on finding better search strategies than in 1991 in the 2000 full database, both through considering additional terms and using more sophisticated strategies.

Conclusion

Methodologic MEDLINE search filters developed in 1991 generally performed at least as well when searching in the publishing year 2000. Until better strategies are devised, users of MEDLINE features such as Clinical Queries can be confident of their enduring performance.

References

[1] de Solla Price D. The development and structure of the biomedical literature. Ch.1 in Warren KS, ed. *Coping with the Biomedical Literature.* New York: Praeger Publishers, 1981: pp. 3-16.

[2] Haynes RB, Sackett DL, Tugwell P. Problems in handling of clinical and research evidence by medical practitioners. *Arch Intern Med* 1983;143:1971-5.

[3] Martinez JL, Licea Serrato J de D, Jimenez R, Grimes RM. HIV/AIDS practice pattern, knowledge, and education needs among Hispanic clinicians in Texas, USA, and Nuevo Leon, Mexico. *Rev Panam Salud Publica* 1998;4:14-9.

[4] Covell MF, Uman GC, Manning PR. Information needs in office practice: are they being met? *Ann Intern Med* 1985;103:596-9.

[5] Balas EA, Stockham MG, Mitchell JA, Sievert ME, Ewigman BG, and Boren SA. In search of controlled evidence for health care quality improvement. *J Med Syst* 1997;21:21-32.

[6] 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.

[7] Ingui BJ, Rogers MA. Searching for clinical prediction rules in MEDLINE. *J Am Med Inform Assoc* 2001;8:391-7.

[8] 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.

[9] 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.

[10] 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.

[11] Wilczynski NL, McKibbon KA, Haynes RB. Enhancing retrieval of best evidence for health care from bibliographic databases: calibration of the hand search of the literature. Medinfo. 2001;10(Pt 1):390-3.

[12] McKibbon KA, Wilczynski NL, Haynes RB. Where are high quality clinically relevant studies published? International Congress on Peer Review in Biomedical Publication. Barcelona, Spain. September 14, 2001.

Acknowledgments

This research was funded by the National Library of Medicine, USA. The Hedges Team includes Angela Eady, Brian Haynes, Susan Marks, Ann McKibbon, Doug Morgan, Cindy Walker-Dilks, Nancy Wilczynski, and Sharon Wong.