

What is already known on this topic

Some people believe that participation in a randomised controlled trial (RCT) increases a patient's risk of a bad outcome

Some people claim that the results of RCTs are not applicable to usual clinical practice

What this study adds

Participants in RCTs had similar outcomes to comparable patients who received the same or similar treatment outside the trial

The results of RCTs are therefore applicable to comparable patients in usual clinical practice

Funding: Norwegian Health Services Research Centre, McMaster University, and the Nuffield Trust.

Competing interests: None declared.

Ethical approval: Not required.

- 1 Stiller CA. Centralised treatment, entry to trials and survival. *Br J Cancer* 1994;70:352-62.
- 2 Braunholtz DA, Edwards SJL, Lilford RJ. Are randomized clinical trials good for us (in the short term)? Evidence for a "trial effect." *J Clin Epidemiol* 2001;54:217-24.

- 3 Emergency Care Research Institute 2002. Patients' reasons for participation in clinical trials and effect of trial participation on patient outcomes. www.ecri.org/Patient_Information/Patient_Reference_Guide/evidence.pdf (accessed Nov 2004).
- 4 Peppercorn JM, Weeks JC, Cook EFC, Joffe S. Comparison of outcomes in cancer patients treated within and outside clinical trials: conceptual framework and structured review. *Lancet* 2004;363:263-70.
- 5 Rothwell PM. External validity of randomised controlled trials: "To whom do the results of this trial apply?" *Lancet* 2005;365:82-93.
- 6 Higgins JP, Thompson SG. Quantifying heterogeneity in meta-analysis. *Stat Med* 2002;21:1539-58.
- 7 Mahon J, Laupacis A, Donner A, Wood T. Randomised study of n of 1 trials versus standard practice. *BMJ* 1996;312:1069-74.
- 8 Mahon JL, Laupacis A, Hodder RV, McKim DA, Paterson NAM, Wood TE, et al. Theophylline for irreversible chronic airflow limitation. A randomized study comparing n of 1 trials to standard practice. *Chest* 1999;115:38-48.
- 9 Dahan R, Caulin C, Figea L, Kanis JA, Caulin F, Segrestaa JM. Does informed consent influence therapeutic outcome? A clinical trial of the hypnotic activity of placebo in patients admitted to hospital. *BMJ* 1986;293:363-4.
- 10 Cooper KG, Grant AM, Garratt AM. The impact of using a partially randomised patient preference design when evaluating alternative managements for heavy menstrual bleeding. *Br J Obstet Gynaecol* 1997;104:1367-73.
- 11 Bergmann JF, Chassany O, Gandiol J, Deblois P, Kanis JA, Segrestaa JM, et al. A randomised clinical trial of the effect of informed consent on the analgesic activity of placebo and naproxen in cancer pain. *Clin Trials* 1994;29:41-7.
- 12 Kunz R, Vist G, Oxman AD. Randomisation to protect against selection bias in healthcare trials (Cochrane methodology review). In: *Cochrane Library*. Issue 4. Oxford: Update Software, 2002.

(Accepted 31 March 2005)

doi 10.1136/bmj.38447.490463.8F

Optimal search strategies for retrieving scientifically strong studies of treatment from Medline: analytical survey

R Brian Haynes, K Ann McKibbin, Nancy L Wilczynski, Stephen D Walter, Stephen R Werre for the Hedges Team

Abstract

Objective To develop and test optimal Medline search strategies for retrieving sound clinical studies on prevention or treatment of health disorders.

Design Analytical survey.

Data sources 161 clinical journals indexed in Medline for the year 2000.

Main outcome measures Sensitivity, specificity, precision, and accuracy of 4862 unique terms in 18 404 combinations.

Results Only 1587 (24.2%) of 6568 articles on treatment met criteria for testing clinical interventions. Combinations of search terms reached peak sensitivities of 99.3% (95% confidence interval 98.7% to 99.8%) at a specificity of 70.4% (69.8% to 70.9%). Compared with best single terms, best multiple terms increased sensitivity for sound studies by 4.1% (absolute increase), but with substantial loss of specificity (absolute difference 23.7%) when sensitivity was maximised. When terms were combined to maximise specificity, 97.4% (97.3% to 97.6%) was achieved, about the same as that achieved by the best single term (97.6%, 97.4% to 97.7%). The strategies newly reported in this paper outperformed other

validated search strategies except for two strategies that had slightly higher specificity (98.1% and 97.6% v 97.4%) but lower sensitivity (42.0% and 92.8% v 93.1%).

Conclusion New empirical search strategies have been validated to optimise retrieval of articles from Medline reporting high quality clinical studies on prevention or treatment of health disorders.

Introduction

If large bibliographic databases such as Medline are to be helpful to clinical users, clinicians must be able to retrieve articles that are scientifically sound and directly relevant to the health problem they are trying to solve, yet few clinicians are trained in search techniques. Search filters ("hedges") can improve the retrieval of clinically relevant and scientifically sound study reports from Medline and similar bibliographic databases.¹⁻⁶ Hedges can be created with appropriate disease content terms combined ("ANDed") with

Editorial by Sanders and Del Mar

Health Information Research Unit, McMaster University, Hamilton, ON, Canada L8N 3Z5

R Brian Haynes
chief

Stephen R Werre
research associate

School of Graduate Studies, McMaster University
Nancy L Wilczynski
doctoral candidate

Department of Clinical Epidemiology and Biostatistics, McMaster University
Stephen D Walter
professor

continued over



This is the abridged version of an article that was posted on bmj.com on 13 May 2005: <http://bmj.com/cgi/doi/10.1136/bmj.38446.498542.8F>

BMJ 2005;330:1179-82

Center for
Biomedical
Informatics,
University of
Pittsburgh,
Pittsburgh, PA, USA
K Ann McKibbin
doctoral candidate

Correspondence to:
R B Haynes
bhaynes@
mcmaster.ca

medical subject headings (MeSH), explosions (px), publication types (pt), subheadings (sh), and textwords (tw) that detect research design features indicating methodological rigour for applied healthcare research.

In the early 1990s, our group developed Medline search filters for studies of the cause, course, diagnosis, or treatment of health problems, based on 10 clinical journals.⁷ Here we report improved hedges for retrieving studies on prevention and treatment, developed on a larger number of journals in a more current era than previously reported.⁸

Methods

Our methods are detailed elsewhere.^{9 10} Briefly, research staff hand searched each issue of 161 clinical journals indexed in Medline for the year 2000 to find studies on treatment that met several criteria (see bmj.com). Search strategies were then created and tested for their ability to retrieve articles in Medline that met these criteria while excluding articles that did not.

Table 1 shows the sensitivity, specificity, precision, and accuracy of single term and multiple term Medline search strategies that we determined. The sensitivity for a given strategy is defined as the proportion of articles retrieved that are scientifically sound and clinically relevant (high quality articles); specificity is the proportion of lower quality articles (did not meet criteria) that are not retrieved; precision is the proportion of retrieved articles that meet criteria (equivalent to positive predictive value in diagnostic test terminology); and accuracy is the proportion of all articles that are correctly dealt with by the strategy (articles that met criteria and were retrieved plus articles that did not meet criteria and were not retrieved divided by all articles in the database).

Eventually a small fraction (n=968, 2%) of citations downloaded from Medline could not be matched to the handsearched data. As a conservative approach, unmatched citations that were detected by a given search strategy were included in cell b of the analysis in table 1. Similarly, unmatched citations that were not detected by a search strategy were included in cell d of the table.

Six research assistants completed rigorous calibration exercises for application of methodological criteria to articles to determine if the article was methodologically sound (see bmj.com).

To construct a comprehensive set of possible search terms, we listed MeSH terms and textwords related to study criteria and then sought input from clinicians and librarians through several means (see bmj.com). We compiled a list of 4862 unique terms (data not shown). All terms were tested using the Ovid searching system.

Manual ratings of articles were recorded on data collection forms along with bibliographic information and database specific unique identifiers. Each journal title was searched in Medline for 2000, and the full Medline records were captured for all articles. Medline data were then linked with the manual data.

Testing strategies

We randomly divided treatment and prevention articles that met criteria in the manual review database into development and validation datasets (60% and 40%). Sensitivity, specificity, precision, and accuracy were calculated for each term in the development subset and then validated in the rest of the database. For a given purpose category, we incorporated individual search terms with sensitivity greater than 25% and specificity greater than 75% into the development of search strategies that included a combination of two or more terms. All combinations of terms used the boolean OR.

For the development of multiple term search strategies to either optimise sensitivity or specificity, we tested all two term search strategies with sensitivity at least 75% and specificity at least 50%. For optimising accuracy, two term search strategies with accuracy greater than 75% were considered for multiple term development. Overall, we tested 18 404 multiple term search strategies. Search strategies were also developed that optimised combined sensitivity and specificity (by keeping the absolute difference between sensitivity and specificity less than 1%, if possible).

To attempt to increase specificity without compromising sensitivity, we used terms with low sensitivity but appreciable specificity to NOT out citations. We also used logistic regression analysis models that included terms in a stepwise manner and also NOTed out terms with a regression coefficient less than -2.0.

We compared strategies that maximised each of sensitivity, specificity, precision, and accuracy for both development and validation datasets with 19 previously published strategies. We chose strategies that had been tested against an ideal method such as a hand search of the published literature and for which most Medline records were from 1990 forward, to reflect major changes in the classification of clinical trials by the US National Library of Medicine. Six papers¹⁻⁶ and one library website¹¹ provided a total of 19 strategies to test, including the strategy advocated by the Cochrane Collaboration.¹

Results

We included 49 028 articles in the analysis; 6568 articles (13.4%) were classified as original studies evaluating a treatment, of which 1587 (24.2%) met our methodological criteria. Overall, 3807 of 4862 proposed unique terms retrieved citations from Medline that could be used in term assessment. The development and validation datasets for assessing retrieval strategies included articles that passed and did not pass treatment criteria (930 and 29 397 articles, respectively, for the development dataset; 657 and 19 631 articles for the validation dataset). The validation dataset provided differences in performance that were statistically significant in only three of 36 comparisons.

Table 1 Formula for calculating sensitivity, specificity, precision, and accuracy of Medline searches for detecting sound clinical studies

Search terms	Manual review	
	Meets criteria	Does not meet criteria
Detected	a	b
Not detected	c	d
	a+c	b+d

Sensitivity=a/(a+c); precision=a/(a+b); specificity=d/(b+d); accuracy=(a+d)/(a+b+c+d). All articles classified during manual review of literature=(a+b+c+d).

Table 2 Comparison of strategies from 1991 with newly developed strategies from 2000, compiled using 2000 data

Year	Approach	Strategy in Ovid format	Sensitivity (%)	Specificity (%)	Precision (%)	Accuracy (%)
1991	Maximise sensitivity	Randomized controlled trial.pt. OR drug therapy.fs. OR therapeutic use.fs. OR random:.tw.	98.3	79.4	13.4	78.6
2000	Maximise sensitivity	Clinical trial.pt. OR random:.mp. OR therapeutic use.fs.	98.9	79.7	13.8	80.3
1991	Maximise specificity	Double.tw. AND blind.tw. OR placebo:.tw.	42.3	98.1	42.2	96.3
2000	Maximise specificity	Randomized controlled trial.mp. OR randomized controlled trial.pt.	93.1	97.4	54.4	97.3

pt=publication type; fs=floating subheading; :=truncation; tw=textword (word or phrase appears in title or abstract); mp=multiple posting (term appears in title, abstract, or MeSH heading).

The operating characteristics for the single terms with the highest sensitivity and the highest specificity showed that the accuracy is driven by the specificity. The term with the best accuracy when keeping sensitivity more than 50% was “randomized controlled trial.pt.” (see *bmj.com*). The single term that yielded the best precision while keeping sensitivity more than 50% was also “randomized controlled trial.pt.”, and this strategy also gave the optimal balance of sensitivity and specificity.

Strategies combining up to three terms that yielded the highest sensitivity, specificity, and accuracy are shown on *bmj.com*. Some two term strategies outperformed one term and multiple term strategies. The top three search strategies optimising the trade-off between sensitivity and specificity are on *bmj.com*, as are the best combination of terms for optimising the trade-off between sensitivity and specificity when using the boolean NOT to eliminate terms with the lowest sensitivity. Statistically insignificant differences were shown when citations retrieved by the three terms “review tutorial.pt.”, “review academic.pt.”, and “selection criteri:.tw.” were removed from the strategy that optimised sensitivity and specificity.

After the two term and three term computations, search strategies with sensitivity more than 50% and specificity more than 95% were further evaluated by adding search terms selected by using logistic regression modelling. Initially, candidate terms for addition to the base strategy were ordered with the most significant first, using stepwise logistic regression, and then added to the model sequentially. The resulting logistic function (data not shown) determined the association between the predicted probabilities and observed responses. We selected the best one term, two term, three term, and four term strategies. Two had already been evaluated (“randomized controlled trial.mp.” OR “randomized controlled trial.pt.” and “randomized controlled trial.mp.” OR “randomized controlled trial.pt.” OR “double-blind:.tw.”). The other two strategies are listed on *bmj.com*; both had high performance. We next took the 13 terms that had regression coefficients less than -2.0 (“predict.tw.”, “predict.mp.”, “economic.tw.”, “economic.mp.”, “survey.tw.”, “survey.mp.”, “hospital mortality.mp.tw.”, “hospital mortalit:.mp.”, “accuracy:.tw.”, “accuracy.tw.”, “accuracy.mp.”, “explode bias (epidemiology)”, and “longitudinal.tw.”) and NOTed these terms out of the four term search strategy to determine if these terms would improve the operating characteristic values (see *bmj.com*). We found a small but insignificant decrease in sensitivity and increases in specificity, precision, and accuracy.

We compared our best strategies for maximising sensitivity (sensitivity $>99\%$ and specificity $>70\%$) and for maximising specificity while maintaining a high sensitivity (sensitivity $>94\%$ and specificity $>97\%$). To ascertain if the less sensitive strategy (which had a much greater specificity) would miss important articles, we assessed the methodologically sound articles that had not been retrieved by the less sensitive strategy, using studies from the *BMJ*, *JAMA*, *Lancet*, and *New England Journal of Medicine*. In total, 32 articles were missed by the less sensitive search, of which four were from these journals. A practising clinician with training in methods for health research found only one of the four articles to be of substantial clinical importance.¹² The indexing terms for this randomised controlled trial did not include “randomized controlled trial.(pt)”. We used our data to test 19 published strategies^{1-6 11} and we compared these with the best strategies for optimising sensitivity and specificity. The published strategies had a sensitivity range of 1.3% to 98.8% on the basis of handsearched data. All of these were lower than our best sensitivity of 99.3%. The specificities for the published strategies ranged from 63.3% to 96.6%.

Discussion

We have presented a variety of tested search strategies for retrieval of high quality and clinically ready studies of treatments. This research updates our previous hedges published in 1991,⁷ calibrated using 10 medical journals. When these 1991 strategies were tested in the 2000 database, the performance of the 2000 strategies was slightly better for the strategy that maximised sensitivity and considerably better for the strategy that maximised specificity (table 2).

No one search strategy will perform perfectly, for several reasons. Indexing inconsistencies affect retrievals.¹² Indexing terms and methods are modified over time and few changes are implemented retrospectively. Indexers also choose only a small number of terms for each item they index, and many of these terms have similar meanings. Methods and their naming also change over time, and authors may also be imprecise in their description of methods and results, affecting retrievals that are based on textwords in the titles and abstracts. The model we used for testing search strategies defines the constant features of these strategies, their sensitivity and specificity, and these strategies can be expected to perform the same way in the entire Medline database, as shown by their performance in the validation database in our study and by the robustness of the 1991 strategies when retested in our much

What is already known on this topic

Many clinicians and researchers conduct Medline searches independently but lack skills to do this well

A barrier to searching for evidence in Medline is the difficulty selecting an optimal strategy to search for information

What this study adds

Special Medline search strategies were developed and tested that retrieve 99% of scientifically sound therapy articles

Clinicians can use the most specific search when looking for a few sound articles on a topic

Researchers can use the most sensitive search when carrying out a comprehensive search for trials for their systematic reviews

larger 2000 database.⁸ The precision of searches, however, depends on the concentration of relevant articles in the database. We selected clinical journals to calibrate the search strategies, but Medline contains many non-clinical journals. Thus, the concentration of high quality treatment studies will be less in the full Medline database, and the precision of searches will be less accordingly.

Searchers who want retrieval with little non-relevant material can choose strategies with high specificity. For those interested in comprehensive retrievals, strategies with higher sensitivity will be more appropriate. The most effective way to harness these strategies is to have them embedded within searching systems. The most sensitive and most specific search strategies reported here have been implemented in the Clinical Queries search screen (www.ncbi.nlm.nih.gov/entrez/query/static/clinical.html) and by Ovid Technologies (www.ovid.com), and the optimal strategy has been added to Skolar (www.skolar.com).

The Hedges Team includes Angela Eady, Brian Haynes, Susan Marks, Ann McKibbin, Doug Morgan, Cindy Walker-Dilks, Stephen Walter, Stephen Werre, Nancy Wilczynski, and Sharon Wong, all at McMaster University Faculty of Health Sciences.

Contributors: See bmj.com

Funding: National Institutes of Health (grant RO1 LM06866-01).

Competing interests: None declared.

Ethical approval: Not required.

- 1 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.
- 2 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.
- 3 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.
- 4 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.
- 5 Dumbriqne HB, Esquivel JF, Jones JS. Assessment of Medline search strategies for randomized controlled trials in prosthodontics. *J Prosthodont* 2000;9:8-13.
- 6 Jadad AR, McQuay HJ. A high-yield strategy to identify randomized controlled trials for systematic reviews. *Online J Curr Clin Trials* 1993;No 33.
- 7 Haynes RB, Wilczynski N, McKibbin KA, Walker CJ, Sinclair JC. Developing optimal search strategies for detecting clinical sound studies in Medline. *J Am Med Assoc* 1994;271:447-58.
- 8 Wilczynski NL, Haynes RB. Robustness of empirical search strategies for clinical content in Medline. *Proc AMIA Symp* 2002;904-8.
- 9 Haynes RB, Wilczynski NC for the Hedges Team. Optimal search strategies for retrieving scientifically strong studies of diagnosis from Medline. *BMJ* 2004;328:1040-2.
- 10 Montori VM, Wilczynski NL, Morgan D, Haynes RB for the Hedges Team. Optimal search strategies for retrieving systematic reviews from Medline: an analytical survey. *BMJ* 2005;330:68-73.
- 11 University of Rochester Medical Center. Edward G. Miner Library. Evidence-based filters for Ovid Medline. www.urmc.rochester.edu/Miner/links/eBMLinks.html#TOOLS (accessed 22 May 2003).
- 12 Julien JP, Bijker N, Fentiman IS, Peterse JL, Delledonne V, Rouanet P, et al. Radiotherapy in breast-conserving treatment for ductal carcinoma in situ: first results of the EORTC randomised phase III trial 10853. EORTC Breast Cancer Cooperative Group and EORTC Radiotherapy Group. *Lancet* 2000;355:528-33.

(Accepted 31 March 2005)

doi 10.1136/bmj.38446.498542.8F

Medical dress code

We sometimes have to help our students to understand that what they wear influences the way they are seen by patients and colleagues, so affecting their ability to do the job. We may find that we are hesitant to point this out to them nowadays (even when faced, for example, with a bare midriff) for fear of being thought a fuddy-duddy. It is sometimes said that this laxity of dress is a modern phenomenon. I'm not so sure.

I recall going for my anatomy viva in June 1964, at the end of my first year at Cambridge. I had heard that one was expected to wear a gown on these occasions, so I put on my gown over my usual uniform. The examiner was a Dr Bull, an elderly (or so he seemed to me) anatomy lecturer of rather Victorian appearance, with mutton chop whiskers and beetling eyebrows. We went through the viva. The only question I can still remember was being asked to identify a beautiful dissection of the superior and inferior hemiazygos veins. This was so much in contrast with my own sad efforts in the dissecting room that I could not help expressing my admiration. To be honest, I think I might also have

been trying to flatter Dr Bull, in the hope that he had done the dissection himself. At any rate, when the time was up Dr Bull drew himself up, looked at me, and said, "Well, Rushton, you've passed."

"Thank you, Sir."

"But in future please remember, when you come to examinations here, we expect you to wear a tie." He looked at my bare neck.

"A jacket." He looked at my sweater.

"Trousers." He looked at my jeans.

"And shoes." He looked at my sandals.

"Yes, Sir." Mortified, I fled.

The rebuke had been given gently, but, as you see, I do remember it.

David N Rushton *consultant in rehabilitation, Frank Cooksey Rehabilitation Unit, King's College Hospital, London*
(david.rushton@kingsch.nhs.uk)