Reasons for the Loss of Sensitivity and Specificity of Methodologic MeSH Terms and Textwords in MEDLINE

NL Wilczynski, CJ Walker, KA McKibbon, RB Haynes
Health Information Research Unit,
Dept. of Clinical Epidemiology & Biostatistics,
McMaster University, HSC, Room 3H7,
1200 Main St W, Hamilton, Ont, Canada L8N 3Z5
(905)525-9140 x22311, FAX 905-546-0401, E-MAIL WILCZYN@McMASTER.CA

ABSTRACT

Objective: To determine the reasons for the loss of sensitivity and specificity of methodologic MeSH terms and textwords in MEDLINE for identifying sound clinical studies of the etiology, prognosis, diagnosis, prevention, or treatment of disorders in adult general medicine.

Design: Analytic survey of the information retrieval properties of methodologic MeSH terms and textwords selected to detect studies meeting basic methodologic criteria for direct clinical use in general adult medicine.

Measures: Frequency of non-use and misuse of relevant methodologic MeSH terms and textwords among studies meeting and not meeting the basic criteria for clinical practice as determined by the manual review (the gold standard) of all articles in 10 internal and general medicine journals for 1986 and 1991.

Results: Loss of sensitivity due to the non-use of relevant methodologic terms among articles meeting basic methodologic criteria was more pronounced in the areas of diagnosis, prognosis, and etiology than treatment in 1991 and 1986. The use of relevant methodologic terms has improved from 1986 to 1991 in all areas except prognosis. Loss of specificity due to the use of relevant methodologic terms among articles not meeting basic methodologic criteria occurred most frequently in the areas of treatment and etiology.

Conclusions: Although the appropriate use of methodologic MeSH and textwords has improved from 1986 to 1991 among studies meeting basic methodologic criteria for direct clinical use in general adult medicine much improvement is still needed in the areas of diagnosis, prognosis, and etiology. Improvement is needed in assigning the relevant methodologic index terms to studies that meet the methods criteria and in having the authors use the relevant methodologic textwords in the title or abstract. Some improvement is also needed in not using methodologic terms when the study clearly does not meet the methods criteria.

INTRODUCTION

It is important for clinical end users of MEDLINE to be able to retrieve articles that are both scientifically sound and directly relevant to clinical practice. MEDLINE, however, is a general purpose biomedical research literature database, with only a small proportion of articles reporting evidence that can be directly applied in clinical practice. Past research [1-3] has shown that "methodologic search filters" can improve the detection of studies of high quality for clinical practice. A methodologic search filter is a search term or terms (such as 'random allocation' for sound studies of medical intervention) that select studies that are at the most advanced stages of testing for clinical application. The performance of methodologic Medical Subject Headings (MeSH) and textwords vary greatly in MEDLINE and change from year to year [1-3]. In this paper, we report on the reasons for the loss of sensitivity and specificity of individual MeSH terms and textwords for identifying studies meeting basic methodologic criteria on the etiology, prognosis, diagnosis, prevention and treatment of disorders in general adult medicine. Our results are of most interest to indexers, 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 methods of this study has been previously reported in detail [3]. Briefly, to evaluate MEDLINE strategies designed to retrieve studies meeting basic methodologic criteria for clinical practice, terms related to research design features were run as search strategies and treated as "diagnostic tests" for sound studies as determined by the manual review of the literature, treated as the "gold standard". Borrowing from the concepts of diagnostic test evaluation and library science, the sensitivity, specificity, and precision of MEDLINE searches were determined as shown in Table 1. For example, the sensitivity of each MEDLINE search

0195-4210/95/$5.00

term or phrase was calculated as the proportion of relevant, sound citations detected by the search strategy.

Table 1
Formula for Calculating the Sensitivity, Specificity, and Precision of MEDLINE Searches

		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)

To determine the reasons for the loss of sensitivity of individual search terms we calculated the extent of **non-use** of relevant methodologic textwords by the authors of the article, and MeSH by National Library of Medicine (NLM) indexers among studies meeting basic criteria for clinical practice as determined by the manual review of the literature. To determine the reasons for the loss of specificity of individual search terms we calculated the extent of **use** of relevant methodologic textwords and MeSH terms among studies that did not meet basic criteria for clinical practice as determined by the manual review of the literature.

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.

Articles were classified for format, interest, purpose, and methodologic rigor. The format categories and their corresponding definitions are shown in Table 2. For example, an original study was defined as any full-text article in which the investigators made firsthand observations. Items excluded from classification included bannered letters to the editor, book reviews, announcements, policy watch, editorials, brief clinical observations, correspondence, news, obituaries, post-graduate and continuing-education forums, and notices.

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 of 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 alternative, for the purpose of defining optimal methods of management.
Case report	An original study involving less that ten subjects

To be considered of interest to 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.

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. For example, an article was classified as etiology if the content pertained directly to causation of a disease or condition.

Table 3
The Purpose Categories and Their Corresponding Definitions Used to Classify Journal Articles

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.

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. For example, an article classified as prognosis met the basic methodologic criterion if there was a cohort of subjects who had the disease in question at baseline without the outcome of interest.

Table 4
Key Methodologic Criterion, According to Purpose Category, Used to Determine the Methodologic Rigor of Journal Articles

Purpose	Key Methodologic Criterion
Etiology	Formal control group: random or quasi-random allocation of participants to exposure 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).

The manual review of the literature served as the "gold standard" against which MEDLINE search terms (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. 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 (complete list of terms can be found in reference 3).

TESTING STRATEGIES

The sensitivity and specificity of all methods terms were calculated. To determine the reasons for loss of sensitivity and specificity we calculated the use and misuse of relevant methodologic terms. For etiology, the MeSH terms tested are exp case control studies, and exp cohort studies, and the textwords are cohort, case and control:, and case and comparison. For prognosis, the MeSH term tested is exp cohort studies, and the textwords are inception and cohort. For diagnosis, the MeSH term is exp sensitivity and specificity, and the textwords are sensitivity, specificity, and likelihood and ratio:. For treatment, the MeSH terms are random allocation, exp clinical trials, clinical trials, randomized controlled trials, clinical trial (pt), randomized controlled trial (pt), and the textwords are random:, and controlled and trial:. Not all terms were tested in both 1991 and 1986 as some terms were only available in one year.

RESULTS

To determine the reasons for loss of sensitivity the extent of **non-use** of relevant methodologic MeSH terms and textwords among studies that **met** the basic criteria for clinical practice as determined by the manual review of the literature was determined as shown in Table 5 for 1991 and Table 6 for 1986. For example, among 111 articles meeting the basic methodologic criterion for clinical practice in the area of diagnosis in 1991, 48 (43%) did not have the textword sensitivity in the title or abstract.

Table 5
Frequency of Non-use of Relevant Methodologic Terms Among Articles Meeting Basic Methodologic Criteria in 1991

TERMS	FREQUENCY OF NON-USE No. of articles (%)
Etiology (201 articles met methodologic criterion)	
Exp case control studies or exp cohort studies	79 (39%)
Cohort (tw)	152 (76%)
Case and control: (tw	154 (77%)
Case and comparison (tw)	193 (96%)
Prognosis (133 articles met methodologic criterion)	
Exp cohort studies	53 (40%)
Inception and cohort (tw)	132 (99%)
Diagnosis (111 articles met methodologic criterion)	
Exp sensitivity and specificity	55 (50%)
Sensitivity (tw)	48 (43%)
Specificity (tw)	51 (46%)
Likelihood and ratio: (tw)	110 (99%)
Treatment (281 articles met methodologic criterion)*	
Random allocation	273 (97%)
Clinical trial (pt)	21 (7%)
Randomized controlled trial (pt)	36 (13%)
Random: (tw)	31 (11%)
Controlled and trial: (tw)	190 (68%)

* 3 terms were never used: Exp clinical trials, Clinical trials, and Randomized controlled trials.

To determine the reasons for loss of sensitivity we examined in detail articles describing randomized clinical trials (as assessed by the manual review of the literature) that were not indexed with the term clinical trial (pt). This occurred for 21 articles (7%). Each article was retrieved to determine where in the title, abstract, or text the author indicated that the allocation of patients to treatment groups was random. 11 of the 21 articles (52%) had the word random: in the abstract, 9 (43%) had the word random: only in the methods section of the article, and 1 article was incorrectly identified as a randomized controlled trial by the manual review of the literature.

Table 6
Frequency of Non-use of Relevant Methodologic Terms Among Articles Meeting Basic Methodologic Criteria in 1986

TERMS	FREQUENCY OF NON-USE No. of articles (%)
Etiology (155 articles met methodologic criterion)*	
Exp case control studies or exp cohort studies	109 (70%)
Cohort (tw)	141 (91%)
Case and control: (tw)	134 (86%)
Prognosis (106 articles met methodologic criterion)	
Exp cohort studies	37 (35%)
Diagnosis (92 articles met methodologic criterion)	
Exp sensitivity and specificity	88 (96%)
Sensitivity (tw)	52 (57%)
Specificity (tw)	55 (60%)
Likelihood and ratio: (tw)	90 (98%)
Treatment (270 articles met methodologic criterion)	
Random allocation	81 (30%)
Clinical trial (pt)	80 (30%)
Clinical trials	80 (30%)
Random: (tw)	48 (18%)
Controlled and trial: (tw)	192 (72%)

* 1 term was never used: Case and comparison (tw).

When comparing the frequency of non-use of relevant methodologic terms among studies meeting the basic methodologic criteria in 1991 with 1986, there was an improvement in appropriate use of terms in all areas (i.e., etiology, diagnosis, and treatment) except prognosis.

To determine the reasons for loss of specificity the extent of use of relevant methodologic MeSH terms and textwords among studies that **did not** meet the basic criteria for clinical practice as determined by the manual review of the literature was determined as shown in Table 7 for 1991 and Table 8 for 1986. For example, among 181 articles not meeting the basic methodologic criterion for clinical practice in the area of diagnosis in 1991, 9 (5%) had the textword sensitivity in the title of abstract.

A more detailed review of the article resulted when terms were used for studies not meeting the methodologic criterion in the area of treatment. The article was reviewed manually and the design of the study was determined as shown in tables 7 and 8. For example, 42 of the 53 articles that were indexed with clinical trial (pt) in 1991 were case series, 7 were case reports, 3 were trials with contemporaneous controls, and 1 was a trial with historical controls from the same center.

Table 7
Frequency of Use of Relevant Methodologic Terms Among Articles Not Meeting Basic Methodologic Criteria in 1991

TERMS	FREQUENCY OF USE No. of articles (%)
Etiology (209 articles did not meet methodologic criterion)*	
Exp case control studies	16 (8%)
Exp cohort studies	31 (15%)
Cohort (tw)	10 (5%)
Case and control: (tw)	15 (7%)
Prognosis (36 articles did not meet methodologic criterion)†	
Exp cohort studies	2 (6%)
Diagnosis (181 articles did not meet methodologic criterion)‡	
Exp sensitivity and specificity	10 (6%)
Sensitivity (tw)	9 (5%)
Specificity (tw)	9 (5%)
Treatment (358 articles did not meet methodologic criterion)	
Random allocation	1 (1%) 1=CS
Clinical trial (pt)	53 (15%) 42=CS; 7=CR; 3=CC; 1=HCS
Randomized controlled trial (pt)	3 (1%) 2=CS; 1=CR
Random: (tw)	13 (4%) 9=CS; 3=CC; 1=CR
Controlled and trial: (tw)	9 (3%) 5=CS; 2=CR; 1=CC; 1=HCD

* 1 term was never used: Case and comparison (tw).
† 1 term was never used: Inception and cohort: (tw).
‡ 1 term was never used: Likelihood and ratio: (tw).
CS=Case Series; CR=Case Report; HCS=Historical Controls from the Same Center; HDC=Historical Controls from a Different Center; CC=Contemporaneous Controls.

When comparing the frequency of use of relevant methodologic terms among studies not meeting the basic methodologic criteria in 1991 with 1986, there was an improvement in appropriate use of terms in the areas of diagnosis and prognosis. This improvement was not seen in the areas of etiology and treatment.

DISCUSSION

The results of this study show that the appropriate use of MeSH terms and textwords improved among studies meeting basic methodologic

criteria for direct clinical use in the areas of etiology, diagnosis, and treatment when comparing 1991 with 1986. This is not the case, however, for prognosis. Aside from there being a general lack of relevant methods terms to test in the area of prognosis, the 1 MeSH term, exp cohort studies, that could be compared between 1986 and 1991 was less frequently used among articles meeting basic methodologic criteria in 1991.

Table 8
Frequency of Use of Relevant Methodologic Terms Among Articles Not Meeting Basic Methodologic Criteria in 1986

TERMS	FREQUENCY OF USE No. of articles (%)
Etiology (327 articles did not meet methodologic criterion)*	
Exp case control studies	4 (1%)
Exp cohort studies	21 (6%)
Cohort (tw)	2 (1%)
Case and control: (tw)	13 (4%)
Case and comparison (tw)	1 (1%)
Prognosis (30 articles did not meet methodologic criterion)	
Exp cohort studies	3 (10%)
Diagnosis (281 articles did not meet methodologic criterion)*	
Exp sensitivity and specificity	2 (1%)
Sensitivity (tw)	5 (2%)
Specificity (tw)	3 (1%)
Treatment (511 articles did not meet methodologic criterion)	
Random allocation	7 (1%) 3=CS; 2=CC; 1=CR
Clinical trial (pt)	23 (5%) 15=CS; 3=CR; 2=CC; 2=HCS; 1=HCD
Clinical trials	25 (5%) 16=CS; 4=CR; 2=CC; 2=HCS; 1=HCD
Random: (tw)	14 (3%) 5=CS; 4=CC; 2=HCD; 2=CR; 1=HCS
Controlled and trial: (tw)	10 (2%) 5=CS; 2=CR; 2=CC; 1=HCS

* 1 term was never used: Likelihood and ratio: (tw).
CS=Case Series; CR=Case Report; HCS=Historical Controls from the Same Center; HDC=Historical Controls from a Different Center; CC=Contemporaneous Controls.

The loss of sensitivity due to the non-use of relevant methodologic terms was minimal for some of the terms in the area of treatment, e.g. clinical trial (pt). The loss of sensitivity was much more extensive, however, in the area of diagnosis, e.g. exp sensitivity and specificity.

Loss of specificity does not appear to be as big a problem as the loss of sensitivity. The use of relevant methodologic terms among articles not meeting basic methodologic criteria decreased from 1986 to 1991 in the areas of diagnosis and prognosis. However, in the areas of treatment and etiology there was a slight increase in the use of relevant methodologic terms among studies that did not meet the criteria.

Previous studies also have investigated the accuracy of methodologic terms in identifying randomized clinical trials in the area of treatment and have reviewed the reasons for the loss of sensitivity [4-8]. Our results support previous suggestions that identification of randomized controlled reports by indexers might be improved if authors improve titles, abstracts, and lists of keywords. This may also be the case in the areas of diagnosis, prognosis and etiology.

ACKNOWLEDGEMENTS

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

References

[1]. Wilczynski NL, Walker CJ, McKibbon KA, Haynes RB. Assessment of methodologic search filters in MEDLINE. Proc Annu Symp Comp Appl Med Care. 1993;17:601-5.

[2]. Wilczynski NL, Walker CJ, McKibbon KA, Haynes RB. Quantitative comparison of pre-explosions and subheadings with methodologic search terms in MEDLINE. Proc Annu Symp Comp Appl Med Care. 1994;18:905-9.

[3]. Haynes RB, Wilczynski NL, McKibbon KA, Walker CJ, Sinclair JC. Developing optimal search strategies for detecting clinically sound studies in MEDLINE. J Am Med Informatics Assoc. 1994;1:447-58.

[4]. 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.

[5]. 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.

[6]. 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.

[7]. Gotzsche PC, Lange B. Comparison of search strategies for recalling double-blind trials from MEDLINE (abstract). Dan Med Bull. 1991;38:476-8.

[8]. 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).