MEDINFO 2001
V. Patel et al. (Eds)
Amsterdam: IOS Press

Enhancing Retrieval of Best Evidence for Health Care from Bibliographic Databases: Calibration of the Hand Search of the Literature

Nancy L. Wilczynski[a], K. Ann McKibbon[a], R. Brian Haynes[a]

[a]*Health Information Research Unit, McMaster University, Hamilton, Canada*

Abstract

Background: Medical practitioners have unmet information needs. Health care research dissemination suffers from both "supply" and "demand" problems. One possible solution is to develop methodologic search filters ("hedges") to improve the retrieval of clinically relevant and scientifically sound study reports from bibliographic databases. To develop and test such filters a hand search of the literature was required to determine directly which articles should be retrieved, and which not retrieved.
Objective: To determine the extent to which 6 research associates can agree on the classification of articles according to explicit research criteria when hand searching the literature.
Design: Blinded, inter-rater reliability study.
Setting: Health Information Research Unit, McMaster University, Hamilton, Ontario, Canada.
Participants: 6 research associates with extensive training and experience in research methods for health care research.
Main outcome measure: Inter-rater reliability measured using the kappa statistic for multiple raters.
Results: After one year of intensive calibration exercises research staff were able to attain a level of agreement at least 80% greater than that expected by chance (kappa statistic) for all classes of articles.
Conclusion: With extensive training multiple raters are able to attain a high level of agreement when classifying articles in a hand search of the literature.

Keywords:

Observer Variation; Reproducibility of Results; Medicine, Evidence-Based.

Introduction

The conclusion that medical practitioners have unmet information needs is inescapable. 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 (5000) are searchable through electronic databases (eg, MEDLINE), but indexing is course-grained and not very reliable; and retrieval problems for clinical end-users are multiplied by the very low concentration of studies that are new, sound, and ready for application [1], and by the ever increasing number of citations now estimated to be 8000/week entered into MEDLINE. On the demand side, practitioners have difficulties with keeping up-to-date with new advances in health care [2, 3]; most researchable information needs are unmet [4]; and practitioners do not search the medical literature very effectively, their searches lack sensitivity, specificity, and precision [5]. If large electronic bibliographic 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.

One possible solution to these problems is to develop methodologic search filters ("hedges") to improve the retrieval of clinically relevant and scientifically sound study reports from large, general purpose, biomedical research bibliographic databases, such as, MEDLINE, EMBASE, PsycLit, and CINAHL. 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))'. The research to develop this type of search filter was done at McMaster University, Canada, in the early 1990s on a small subset of MEDLINE journals and for 4 types of journal articles [6]. This research is being updated and expanded on using data from the year 2000.

To develop and test such filters in electronic databases a comprehensive list of index terms and textwords that might be used to retrieve such studies were compiled. These terms and combination of terms will be treated as 'diagnostic tests' and a hand search of the literature will be treated as the 'gold standard'. This paper reports on the methods and

results of the calibration of research staff in order to conduct the hand search of the literature.

Materials and Methods

The Health Information Research Unit at McMaster University in Hamilton, Canada prepares 4 evidence-based medical journals, *ACP Journal Club, Evidence-Based Medicine, Evidence-Based Nursing, and Evidence-Based Mental Health.* These journals are secondary publications designed to help keep health care providers up-to-date with advances in the literature. To produce these journals 6 research associates review 170 journal titles on an ongoing basis and apply methodologic criteria to each item in each issue to determine if the article is potentially eligible for inclusion in the evidence-based publications. Data to develop the search filters is being collected using these same journals. Data collection was expanded upon to accommodate both the requirements of the evidence-based publications and for the development of the search filters. Data collection began in the year 2000 but calibration of the research staff was required before this collection. The calibration exercises and reliability tests took place in 1999.

Research staff were calibrated for collection of the following data: format of the article (Table 1), interest to human health care (Table 2), type of data presentation in review articles (Table 3), age of participants in the study (Table 4), purpose of the article (ie, what question(s) are the investigators addressing) (Table 5), and methodologic rigor for each of the purpose categories except for "qualitative" and "something else" (Table 6).

Table 1 – Format categories

Format type	Definition
Original study	Any full text article in which the authors report first-hand observations.
Review article	Any full text article that is bannered 'review, overview, or meta-analysis' in the title or in a section heading, or it is indicated in the text of the article that the intention was to review, summarize, or highlight the literature on a particular topic.
General article	A general or philosophical discussion of a topic without original observation and without a statement that the purpose was to review a body of knowledge.
Case report	An original study or report that presents only individualized data.

Table 2 – Interest to human health care

Of interest	Definition
Yes	Concerned with the understanding of health care in humans; anything that will have an effect on the patient/subject.
No	Not concerned with the understanding of health care in humans; anything that will not have an effect on the patient/subject (eg, studies that describe the normal development of people; basic science; gender and equality studies in the health profession; or studies looking at research methodology issues).

Table 3 – Categories of data presentation in review articles

Type of data presentation	Definition
Individual patient data	Individual patient data was used in a meta-analysis.
Meta-analysis	The reported summary data were pooled from relevant studies.
Overview	A general discussion of the reviewed studies with no attempt to quantitatively combine the results.

Table 4 – Age categories of ≥ 50% of participants

Category	Definition
Fetus	Fetus
Newborn	Birth to 1 month
Infant	> 1 month to < 24 months
Preschool	2 years to < 6 years
Child	6 years to < 13 years
Adolescent	13 years to < 19 years
Adult	19 years to < 45 years
Middle age	45 years to < 65 years
Aged	65 years to < 80 years
Aged 80	≥ 80 years
ND	Non-discernible

Prior to the first inter-rater reliability test research staff met to develop the data collection form and to develop a document outlining the coding instructions and category definitions using examples from the 1999 literature.

The first inter-rater reliability test was conducted after the two-month development period. Results from this testing was used to refine definitions and educate the research staff. Two subsequent testings were strategically scheduled over the course of 1999 with the objective of reducing ambiguities in the definitions used to classified articles.

Articles to be used in the reliability tests were chosen at random from the 170 titles that are to be used to develop the search filters, by someone not otherwise involved in the study. The articles were packaged for the research associates with the data collection forms and the instructions document. Each research associate independently and blindly reviewed each article and recorded their classifications on the data collection forms.

Data were analyzed using PC-agree (written by Richard Cook and maintained by Dr. Stephen Walter, McMaster University, Hamilton, Canada). The level of agreement

Table 5 – Purpose Categories

Purpose Type	Definition
Etiology	Content pertains directly to determining if there is an association between an exposure and a disease or condition. The question is "What causes people to get a disease or condition?"
Prognosis	Content pertains directly to the prediction of the clinical course or the natural history of a disease or condition with the disease or condition existing at the beginning of the study.
Diagnosis	Content pertains directly to using a tool to arrive at a diagnosis of a disease or condition.
Treatment	Content pertains directly to an intervention for therapy (including adverse effects studies), prevention, rehabilitation, quality improvement, or continuing medical education.
Economics	Content pertains directly to the economics of a health care issue.
Clinical Prediction Guide	Content pertains directly to the prediction of some aspect of a disease or condition.
Qualitative	Content relates to how people feel or experience certain situations, and data collection methods and analyses are appropriate for qualitative data.
Something Else	Content of the study does not fit any of the above definitions.

beyond that expected by chance was calculated for each type of article classification (ie, format, interest, review data presentation, age, purpose, and methodologic rigor). The sample size required to obtain a minimally acceptable kappa of 0.80 with a power of 90% and with a two-tailed testing at a 5% alpha level was a total of 69 articles. This sample size allows for multiple comparisons to determine if the classification of 1 of the 8 purpose categories is out of line.

Results

Three reliability tests were conducted over the course of 1999. The goal was to attain kappas of ≥ 0.80 for all classifications. Kappas initially ranged as low as 0.54. Meetings involving the research staff revealed differences in interpretation of the definitions. Intensive discussion periods and practice sessions using actual articles were used to hone definitions and thus remove ambiguities.

The fourth and final inter-rater reliability test was conducted approximately 14 months after the process commenced using a sample of 72 articles randomly selected across the 170 journal titles. In calculating the kappa statistic for methodologic rigor raters had to agree on the purpose category for the item to be included in the

Table 6 – Methodologic Rigor

Purpose Category	Methodologic Rigor
Etiology	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.
Prognosis	Inception cohort of individuals all initially free of the outcome of interest; Follow-up of ≥80% of patients until the occurrence of a major study end point or to the end of the study; Analysis consistent with study design.
Diagnosis	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.
Treatment	Random allocation of participants to comparison groups; Outcome assessment of at least 80% of those entering the investigation accounted for in 1 major analysis at any given follow up assessment; Analysis consistent with study design.
Economics	Question is a comparison of alternatives; Alternative services or activities compared on outcomes produced (effectiveness) and resources consumed (costs); Evidence of effectiveness must be from a study of real patients that meets the above-noted criteria for diagnosis, treatment, quality improvement, or a systematic review article; Effectiveness and cost estimates based on individual patient data (micro-economics); Results presented in terms of the incremental or additional costs and outcomes of one intervention over another; Sensitivity analysis if there is uncertainty.
Clinical Prediction Guide	Guide is generated in one or more sets of real patients (training set); Guide is validated in another set of real patients (test set).
Review articles	Statement of the clinical topic; Explicit statement of the inclusion and exclusion criteria; Description of the methods; ≥ 1 article must meet the above noted criteria.

Table 7 – Level of Agreement

Classification	Kappa (95% CI)
Format	0.92 (0.89 to 0.95)
Interest	0.87 (0.89 to 0.96)
Review data presentation	0.93 (0.86 to 1.00)
Age of participants	0.93 (0.84 to 1.00)
Purpose	0.81 (0.79 to 0.84)
Methodologic Rigor	0.89 (0.78 to 0.99)

calculation. A high level of agreement beyond that expected by chance was attained for all classifications (Table 7). Purpose category specific kappas ranged from 0.72 for the category "something else" to 0.97 for the "qualitative" category.

Discussion

The process of calibrating 6 research associates for hand searching the literature occurred over the period of 14 months. After 4 inter-rater reliability tests and intervening discussion and clarification of criteria, a high level of agreement was attained. This level of agreement for 6 raters is much higher than that typically achieved in studies of diagnosis usually involving only 2 raters. For instance, Mino and colleagues found that the level of agreement between 2 independent raters on 3 expressed emotion ratings ranged from 0.40 to 0.80 [7]. Also Hohol and colleagues found that the kappa for the level of agreement between 2 neurologists when assessing disability in multiple sclerosis was 0.80 using Disease Steps and 0.54 when using the Expanded Disability Status Scale [8].

Having achieved such a high level of agreement a reasonable 'gold standard' will be established to test the search filters developed using articles published and classified in the year 2000.

The training of the research staff to achieve this level of agreement was extensive. Multiple training periods were required with practice exercises and detailed discussions. If other individuals were to apply the criteria used in this study a training period would be required.

Conclusion

With extensive training, multiple raters are able to attain a high level of agreement beyond that expected by chance when classifying articles in a hand search of the literature.

Acknowledgments

This research was funded by the National Library of Medicine, USA and by HEAL*Net*, Canada.

The research staff involved in the calibration exercise were Angela Eady, Susan Marks, Ann McKibbon, Cindy Walker-Dilks, Nancy Wilczynski, and Sharon Wong. Administrative support was provided by Nancy Bordignon. Dr. Brian Haynes provided the leadership in the development of the classification definitions.

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, and Tugwell P. Problems in handling of clinical and research evidence by medical practitioners. *Arch Intern Med* 1984;143:1971-5.

[3] Martinex JL, Licea Serrato J De D, Jimenex R, and Brimes 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, and 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, Sievaert 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, Wilczynski N, McKibbon KA, Walker CJ, and Sinclair JC. Developing optimal search strategies for detecting clinically sound studies in MEDLINE. *J Am Med Inform Assoc* 1994;1:447-58.

[7] Mino Y, Inoue S, Shimodera S, and Tanaka S. Evaluation of expressed emotion (EE) status in mood disorders in Japan: inter-rater reliability and characteristics of EE. *Psychiatry Res* 2000;93:221-7.

[8] Hohol MJ, Orav EJ, and Weiner HL. Disease steps in multiple sclerosis: a simple approach to evaluate disease progression. *Neurology 1995*;45:251-5.

Address for correspondence

Nancy Wilczynski, Health Information Research Unit, Rm 3H7, McMaster University, 1200 Main Street West, Hamilton, Ontario L8N 3Z5, Canada. wilczyn@mcmaster.ca.