Health Libraries Review 1999, **16,** 141–150

A model for the evaluation of search queries in medical bibliographic sources

ANITA A. H. VERHOEVEN,* PETER M. BOENDERMAKER,† EDZARD J. BOERMA‡ and BETTY MEYBOOM-DE JONG§
**University Library, PO Box 559, 9700 AN Groningen; †Department of General Practice, Ant. Deusinglaan 4, 9713 AW Groningen; ‡Department of Education, Grote Rozenstraat 38, 9712 TJ Groningen; §Department of General Practice, Ant. Deusinglaan 4, 9713 AW Groningen, the Netherlands*

The aim of this study was to develop a model to evaluate the retrieval quality of search queries performed by Dutch general practitioners using the printed Index Medicus, MEDLINE on CD-ROM, and MEDLINE through GRATEFUL MED.

Four search queries related to general practice were formulated for a continuing medical education course in literature searching. The selected potential relevant citations from the course instructor and the 103 course participants together served as the basic set for the three judges to evaluate for (a) relevance and (b) quality, with the latter based on journal ranking, research design and publication type. Relevant individual citations received a citation quality score from 1 (low) to 4 (high). The overall search quality was expressed in a formula, which included the individual citation quality score of the selected and missed relevant citations, and the number of selected non-relevant citations.

The outcome measures were the number and quality of relevant citations and agreement between the judges.

Out of 864 citations, 139 were assessed as relevant, of which 44 citations received an individual citation quality score of 1, 76 of 2, 19 of 3 and none of 4. The level of agreement between the judges was 68% for the relevant citations, and 88% for the non-relevant citations.

We describe a model for the evaluation of search queries based not only on the relevance, but also on the quality of the citations retrieved. With adaptation, this model could be generalized to other professional users, and to other bibliographic sources.

Introduction

Evidence-based medicine requires new skills of physicians including efficient literature searching.[1, 2] Thus, we have introduced literature searching in our continuing medical education courses for general practitioners (GPs). In this study we describe the development of a gold standard that could be used as a reference point in comparing and evaluating searches in bibliographic sources.

The approach most used to assess effectiveness of searchers and their searching has been by assessing the quantity of articles retrieved that are relevant to the search topic.[3] The quantity of articles retrieved has been measured with recall (or 'sensitivity'), and

Correspondence: Ms Anita A. H. Verhoeven, UB University Library, PO Box 559, 9700 AN Groningen, the Netherlands. E-mail: A.A.H.VERHOEVEN@MED.RUG.NL

precision (or 'specificity', or more correctly 'positive predictive value').[3–6] However, these measures are based only on the number of selected relevant citations, and take no account of the quality of the relevant citations. We can assume that a relevant review article in a journal with a high impact factor will probably have more impact on the user than a case report in a journal with a low impact factor. Therefore, there is a need for a measure based not only on the number of relevant and non-relevant citations found, but also on measures for the ranking of the journal, the research design and the publication type.

In developing such a measure the concept of relevance is important. No consensus exists on its meaning.[7, 8] Harter[9] stated that relevance is multidimensional, bringing together psychological, situational and topical aspects assessed by the person with the information need. Relevance judgements are affected by four categories of factors:[10]

- user's perceptions concerning elements of the citation, such as quality of the journal, style of the title of the article, and status of the author;
- internal context, such as experience with journals and authors, and awareness of published literature;
- external context, such as goal of the search and stage of research—the situational relevance;
- problem context, or the characteristics of the user's information problem, e.g. whether the citation cast a new or different light on the problem.

For the purpose of our model, we decided to use a simple measure of topical relevance, and three measures of the quality of the retrieved citations. This did not include outcome-orientated relevance, which measures the use that is made of the search results.[11]

The aim of this article is to describe the development of a model to evaluate the retrieval quality of search queries on more than quantitative measures, performed by Dutch GPs in medical bibliographic sources.

Methods

The model was developed as part of a 1-day continuing medical education course in literature searching for Dutch GPs. The course was offered from 1994 to 1997, and in total 103 GPs took part. The GPs were randomized blockwise in three groups and we assigned each group to one of the following bibliographic sources: the 1992 printed Index Medicus, the 1992 MEDLINE on CD-ROM (Silverplatter version 3.1), and MEDLINE (citations with publication year 1992) through GRATEFUL MED (version 6.0) provided online by the Karolinska Institute in Sweden. We formulated four search queries on topics with varying complexity related to general practice, namely haemorrhoids, sudden infant death, the telephone, and the gatekeeper (Table 1). The topics of the search queries were collected from discussions with practicing GPs.

After 2 hours of instruction, the GPs each searched one of the three sources for relevant citations. First they retrieved a batch of citations and then decided which to select as relevant for the four search queries.

To evaluate the number and quality of the selected citations we developed a gold

Table 1. The four search queries for a course in literature searching for Dutch family physicians.

1. You have been invited to give a talk on haemorrhoids to a group of colleagues in your town. To collect information, you want to do a literature search using the Index Medicus/MEDLINE on CD-ROM/MEDLINE through GRATEFUL MED of the year 1992.
2. You want to keep up with the latest developments in the prevention of sudden infant death. You want to do a literature search using the Index Medicus/MEDLINE on CD-ROM/MEDLINE through GRATEFUL MED of the year 1992.
3. The editor of the Dutch journal *Huisarts & Wetenschap* (*Family Physician & Medical Science*) has asked you to submit an article on the use of the telephone in the physician's office. You regard his request as a challenge and you decide to do a literature search using the Index Medicus/MEDLINE on CD-ROM/MEDLINE through GRATEFUL MED of the year 1992.
4. You are supervising a family physician trainee in your practice. She wants to discuss the gatekeeper role of the family physician. You want to be well prepared for this discussion so you decide to do a literature search on this topic using the Index Medicus/MEDLINE on CD-ROM/MEDLINE through GRATEFUL MED of the year 1992.

standard. Three judges decided which of the citations were relevant, and then they assessed the quality level of these relevant citations. With this information we could assess the retrieval quality of search queries in the form of an overall search quality score.

The relevance judges were all GPs by training: a professor in general practice (judge A), a trainer of GPs (judge B) and an experienced information specialist (judge C).

Relevance of citations

It was impossible to consider all available citations from the three sources, therefore the judges evaluated all citations selected as relevant by the 103 participants on the course as well as the citations selected by the course instructor. The definition of relevance used by the judges was 'whether the journal article represented by the citation would give relevant information for answering the question that prompted the search', based upon the title, the language and, if available, the abstract and the Medical Subject Headings. Relevant languages were English, German, French and Dutch. A letter was only relevant if it was published in a general practice journal, or in a general medical journal such as the *Lancet* or the *New England Journal of Medicine*. Editorials were considered as articles. Disagreements between the judges were resolved by discussion. We then calculated the number of citations the individual judge's performances had in common with the gold standard of the consensus result in two ways: (i) agreement on the relevant citations (sensitivity), and (ii) agreement on the non-relevant citations (specificity).

Quality assessment of citations

We chose three formal criteria: journal ranking, research design, and publication type.

In the literature, peer review and impact factors have been used for assessing the importance of medical journals.[12] Because the peer review process was not described explicitly in most of the journals it was not considered a workable method for our study. We therefore relied on the impact factor of the journal. The impact factor is a measure of the frequency with which the 'average article' in a journal has been cited in a particular year,[13] and therefore gives a rough indication of relative influence of the journal. In

addition, we used the impact factor for pragmatic reasons. Research institutions all over the world, including those receiving Dutch governmental research funds, and Dutch medical schools, use the impact factor as a measure for the scientific output of researchers.

We used the most recent available data from the *Journal Citation Reports* database (science and social science editions).[14, 15] If the journal appeared in both indexes, the *Science Citation Index* prevailed. Because the Citation Indexes have a strong USA bias, we also used the Dutch list of *Additional Scientific Journals for Health Sciences Research* of the Royal Netherlands Academy of Arts and Sciences.[16] This list was compiled by members of the Dutch medical profession, and it covered 106 journals relevant for Dutch GPs.

When a relevant citation was published in a journal that had an impact factor above the median for its subject category, or if it was published in a journal covered by the Dutch list of *Additional Scientific Journals for Health Sciences Research*, it received one point.

The second quality criterion was research design. The citation received one point if the title, abstract or publication type field contained one of the following components of research design: clinical trial, placebo-controlled study, double-blind study, follow-up or prospective study.

The final quality criterion was publication type. If the citation was a review it received one point, because a review gives an overview of the most recent studies on that special subject. Moreover, it provides many other citations for further study. A citation was counted as a review article if:

- the title contained the word review or overview; or
- the citation in Index Medicus mentioned the number of references (in the Index Medicus this is the standard presentation for all reviews); or
- the citation from the CD-ROM or GRATEFUL MED contained review or meta-analysis in the Publication Type (PT) field.

For each selected relevant citation an individual citation quality score was calculated. This score consisted of the sum of the points, ranging from 1 to 4, according to the following criteria:

1. one point for topical relevance (only relevant citations were included, so every citation received 1 point);
2. one point for journal ranking (above the median impact factor for its subject group, or listing in the Dutch list of *Additional Scientific Journals for Health Sciences Research*);
3. one point for research design (clinical trial, placebo-controlled study, double-blind study, follow-up or prospective study);
4. one point for review.

Overall search quality score

Using the scores of the individual relevant citations, we expressed the overall search retrieval quality in a new quality measure instrument: the search quality score. This consisted of the sum of the scores of the selected, relevant citations, reduced by the sum of the scores of the missed relevant citations, and the number of the selected non-relevant citations. To avoid negative values, we added 100 points to these scores.

Table 2. Number of unique citations retrieved by the course instructor, and the number of additional new unique citations retrieved by the 103 course participants.

	Number of unique citations retrieved by course instructor			Number of additional unique citations retrieved by course participants		
Search query	Total	Relevant	Non-relevant	Total	Relevant	Non-relevant
Haemorrhoids	62	22	40	6	0	6
SIDS	66	31	35	92	23	69
Telephone	142	36	106	33	5	28
Gatekeeper	369	20	349	94	2	92
All search queries	639	109	530	225	30	195

In other words, the search quality score = a − b − c + 100, in which a = total score of the selected relevant citations, b = total score of the missed relevant citations, and c = number of selected non-relevant citations.

Results

Relevance of citations

After we deleted duplicate citations which appeared in more than one source, our list of potential relevant citations consisted of 864 unique citations. From these 864 unique citations, 139 were assessed as relevant by the judges, and 725 as non-relevant. From the 864 unique citations, 639 were selected by the course instructor, and an additional 225 new citations were selected by the 103 course participants. The judges indicated 139 citations as relevant, 109 from the search strategies of the course instructor, and 30 additional citations from the search strategies of the course participants (Table 2).

The judges checked relevance differences between the course participants and the judges, but they found no reason to readjust their decisions.

First, all 17 citations selected by more than 10 course participants but assessed by the judges as not relevant were re-evaluated. Of these 17 citations 10 were assessed again as non-relevant for formal reasons (in a language other than English, German, French or Dutch; a letter in a journal other than a general practice or general medical journal). Seven citations were assessed again as non-relevant because they were too specialized for a GP.

Second, all 13 citations that were relevant according to the judges, but were selected by none of the course participants, were re-evaluated. All of the 13 citations were again assessed as relevant by the judges after careful reconsideration.

An analysis of the agreement between the judges and the consensus result (Table 3) showed that:

- On average, each judge indicated 68% of the relevant citations.
- On average, each judge correctly indicated 88% of the non-relevant citations.

In other words, the sensitivity of the judges' selection process was 68%, and the specificity was 88%.

Table 3. Percentage (and number of) citations the three relevance judges and the consensus result have in common.

	Percentage agreement (*n*)				
Search query	Judge A	Judge B	Judge C	Mean per query	Mean all queries
Relevant citations (n = 139)					
Haemorrhoids	96 (21)	27 (6)	64 (14)	62 (14)	
SIDS	75 (41)	67 (36)	93 (50)	78 (42)	68
Telephone	83 (34)	44 (18)	61 (25)	63 (26)	
Gatekeeper	46 (10)	65 (14)	64 (14)	58 (13)	
Non-relevant citations (n = 725)					
Haemorrhoids	70 (32)	100 (46)	98 (45)	89 (41)	
SIDS	81 (44)	94 (51)	88 (48)	88 (48)	88
Telephone	84 (113)	99 (133)	90 (121)	91 (122)	
Gatekeeper	98 (432)	98 (432)	96 (423)	97 (429)	

Quality assessment of citations

Of the 89 journals in which the relevant articles of our study were published, 58 were covered by the *Science Citation Index* and/or by the *Social Science Citation Index*. Thirty-nine had an impact factor above the median of their subject category, and an additional four were covered by the Dutch list of *Additional Scientific Journals for Health Sciences Research*.

Of the 139 relevant citations, 15 had the described components of research design. Of the 139 citations, 27 referred to a review article. None of the reviews was a systematic review.

Table 4 shows the number of relevant citations with a score from 1 to 4. None of the papers received a point for both study design and publication type.

Overall search quality score

The overall search quality score of a search was calculated from the individual citation quality scores and the number of selected non-relevant citations. To illustrate how the formula worked we give an assessment of two specific searches undertaken by course participants A and B.

Participant A performed a search in *Index Medicus* to find references on haemorrhoids. Participant A found six citations under the Medical Subject Heading 'haemorrhoids', all of

Table 4. Number of relevant citations with an individual citation quality score from 1 to 4.

	Individual citation quality score			
Search query	1	2	3	4
Haemorrhoids	7	11	4	0
SIDS	18	25	11	0
Telephone	12	26	3	0
Gatekeeper	7	14	1	0
All search queries	44	76	19	0

which were relevant according to the judges. The total of the score of the selected citations was 9 (a). The total score of the missed citations was 10 (b). The number of the selected non-relevant citations was 0 (c). Hence, the overall search quality score for participant A was a − b − c + 100; 9 − 10 − 0 + 100 = 99. Participant B performed a search in MEDLINE on CD-ROM to find references on the gatekeeper role of the GP. Participant B retrieved 154 citations with the combination of the Medical Subject Headings 'physicians-family', 'family-practice' and 'referral-and-consultation'. Five citations were selected, but none was relevant. The total score of the selected citations was 0 (a). The total score of the missed citations was 11 (b). The number of the selected non-relevant citations was 5 (c). Hence, the overall search quality score for participant B was 0 − 11 − 5 + 100 = 84.

Discussion

We have described the development of a model to evaluate the retrieval quality of search queries expressed in an overall search quality score. This score included the assessment of retrieved citations by quality as well as by a narrowly defined concept of relevance.

New in this study is that we developed a quality measure to assess the overall search results in addition to the determination of the number of relevant citations. In this context it is related to the evidence-based quality filters developed and described by Haynes,[17] which have been built into PUBMED. However, our model differs at some points. First, our model does not influence the course of a search but uses a way of measuring search quality once the search has been performed. Second, it can be used not only for clinical queries that fit into four disease-related study categories (therapy, diagnosis, aetiology, prognosis), but also for non-clinical queries such as the use of the telephone in the physician's office, and the gatekeeper role of the GP. Third, in addition to research design, our model also includes journal ranking and publication type. Finally, our model not only assesses individual retrieved citations, but also the overall search results, including the missed citations.

The methodology of the study raises several points of discussion.

The relevance assessment was not made by the person who prompted the search query, but by independent judges. Although the search questions were linked to daily practice, they were not formulated and judged by the real end user. However, Janes & McKinney[18] found that in general, judges' relevance scores compared reasonably well with those of the original users. Saracevic[19] pointed to the importance of the judges' subjective expertise, knowledge, and academic and professional training. He found that the more judges knew about a query, the higher was the agreement among judges on relevance judgements and the more stringent the judgements became. The judges in our study had a high level of knowledge, expertise and training in the field of general practice. Furthermore, one judge was experienced in online information retrieval.

Another factor is that the selection was made on title, abstract and controlled vocabulary, which does not always represent the content of the article correctly. We deliberately chose this method, because we wanted to use this model in a continuing

medical education course for GPs, who usually select relevant citations only on the basis of data in the bibliographic source.

Finally, many factors determine the selection process of citations. The judges in our study initially considered the topical relevance and language. At a later stage, the relevant citations were assessed by the journal in which they were published, their research design and their publication type. However, more factors can influence the decision process. In an interview study among clinicians, Sievert and co-workers[20] found that factors beyond relevance that most often influenced the decision process to choose a citation not only pertained to methodological rigor and document types, but also to authors, their institutional affiliations and population studied. Furthermore, users' criteria for relevance could vary as they interacted with the information.[21]

Although the quality of citations is influenced by objective factors such as relevance, journal ranking, research design and publication type, other factors could be considered as well. Factors from the external and problem context of the information need[10] are often personal, subjective factors and therefore difficult to grasp, these include the user's purpose in obtaining information, whether the information adds to the user's knowledge, and the applicability of the obtained knowledge. Su investigated the appropriateness of 20 measures for evaluating interactive information retrieval performance, representing four major evaluation criteria: precision and recall, efficiency, utility and user satisfaction.[6] Recall appeared to be more important than precision to users. Users satisfaction with precision of the search and with completeness of search results were important as well.

The ultimate test of end-user searching is whether it actually improves patient outcomes.[4] Hersh pointed to the outcomes-orientated relevance,[11] namely the impact of relevant information on users and their tasks. Some research has been carried out on the impact of the library and information services on medical decision-making and patient care in hospitals,[22–26] but the impact of information services on medical decision-making and patient care in primary health care is scarce.[27, 28] Our overall search quality score can be used to assess high-quality searches; future research could assess the impact of these high-quality searches on medical decision-making and patient care in primary health care.

When we started the study in 1992, GRATEFUL MED was only available in Europe as a standalone with a modem connection to a host computer at the Karolinska Institute in Stockholm. However, when in 1997 the US National Library of Medicine made GRATEFUL MED freely available on the Internet (http://igm.nlm.nih.gov), the Karolinska Institute stopped offering GRATEFUL MED online services. Although the data and search principles are similar, it is not known whether the two software systems are similar in user friendliness and search results.

In summary, this study describes an overall search quality score that can be used for evaluation of the retrieval quality of search queries by GPs in medical bibliographic sources. Although this model was developed for Dutch GPs who retrieved citations from three specific sources, it could be adapted and generalized to other professional users, and to other bibliographic sources, such as MEDLINE on the Internet (Internet GRATEFUL MED, PUBMED or MEDLINE through another Web interface). As a national factor in our model, the journal list can be replaced by a journal list appropriate to another country.

Acknowledgements

The authors thank the Janivo Foundation for their financial support.

References

1 Evidence-based Medicine Working Group. Evidence-based medicine: a new approach to teaching the practice of medicine. *JAMA* 1992, **268**, 2420–5.
2 Geyman, J. P. Evidence-based medicine in primary care: an overview. *Journal of the American Board of Family Practice* 1998, **11**, 46–56.
3 Hersh, W. R. & Hickam, D. H. How well do physicians use electronic information retrieval systems? A framework for investigation and systematic review. *JAMA* 1998, **280**, 1347–52.
4 McKibbon, K. A. & Walker-Dilks, C. J. The quality and impact of MEDLINE searches performed by end users. *Health Libraries Review* 1995, **12**, 191–200.
5 Su, L. T. Evaluation measures for interactive information retrieval. *Information Processing and Management* 1992, **28**, 503–16.
6 Su, L. T. The relevance of recall and precision in user evaluation. *Journal of the American Society for Information Science* 1994, **45**, 207–17.
7 Schamber, L., Eisenberg, M. B. & Nilan, M. S. A re-examination of relevance: toward a dynamic, situational definition. *Information Processing and Management* 1990, **26**, 755–76.
8 Mizzaro, S. Relevance: the whole history. *Journal of the American Society for Information Science* 1997, **48**, 810–32.
9 Harter, S. P. Psychological relevance and information science. *Journal of the American Society for Information Science* 1992, **43**, 602–15.
10 Park, T. K. The nature of relevance in information retrieval: an empirical study. *Library Quarterly* 1993, **63**, 318–51.
11 Hersh, W. Relevance and retrieved evaluation: perspectives from medicine. *Journal of the American Society for Information Science* 1994, **45**, 201–6.
12 Weller, A. C. Editorial policy and the assessment of quality among medical journals. *Bulletin of the Medical Library Association* 1987, **75**, 310–6.
13 Garfield, E. *SCI Journal Citation Reports; a Bibliometric Analysis of Science Journals in the ISI Database.* Philadelphia: Institute for Scientific Information, 1994: 10.
14 Garfield, E. *SCI Journal Citation Reports; a Bibliometric Analysis of Science Journals in the ISI Database.* Philadelphia: Institute for Scientific Information, 1994: 72–93.
15 Garfield, E. *SSCI Journal Citation Reports; a Bibliometric Analysis of Social Science Journals in the ISI Database.* Philadelphia: Institute for Scientific Information, 1995: 40–6.
16 Medical Committee of the Royal Netherlands Academy of Arts and Sciences, and the Council of Medical Faculties in the Netherlands of the Association of Universities in the Netherlands. Additional scientific journals for health sciences research. In: *Guidelines for provision of information for the Discipline Report on (Bio) Medical and Health Sciences Research 1998.* In Dutch (*Richtlijnen aanlevering gegevens ten behoeve van het Discipline-advies Geneeskunde 1998*). Amsterdam, Utrecht: Medical Committee and Council of Medical Faculties, 1996: 57.
17 Haynes, R. B., Wilczynski, N., McKibbon, K. A., Walker, C. J. & Sinclair, J. C. Developing optimal search strategies for detecting clinically sound studies in MEDLINE. *Journal of the American Medical Informatics Association* 1994, **1**, 447–58.
18 Janes, J. W. & McKinney, R. Relevance judgments of actual users and secondary judges: a comparative study. *Library Quarterly* 1992, **62**, 150–68.
19 Saracevic, T. Relevance: a review of and a framework for the thinking on the notion in information science. *Journal the American Society for Information Science* 1975, **26**, 321–43.
20 Sievert, M. E., McKinin, E. J., Johnson, E. D., Reid, J. C. & Mitchell, J. A. Beyond relevance—characteristics of key papers for clinicians: an exploratory study in an academic setting. *Bulletin of the Medical Library Association* 1996, **84**, 351–8.
21 Quintana, Y. Intelligent medical information filtering. *International Journal of Medical Informatics* 1998, **51**, 197–204.

22 King, D. N. The contribution of hospital library information services to clinical care: a study in eight hospitals. *Bulletin of the Medical Library Association* 1987, **75**, 291–301.

23 Marshall, J. G. The impact of the hospital library on clinical decision making: the Rochester study. *Bulletin of the Medical Library Association* 1992, **80**, 169–78.

24 Lindberg, D. A. B., Siegel, E. R., Rapp, B. A., Wallingford, K. T. & Wilson, S. R. Use of MEDLINE by physicians for clinical problem solving. *JAMA* 1993, **269**, 3124–9.

25 Klein, M. S., Ross, F. V., Adams, D. L. & Gilbert, C. M. Effect of online literature searching on length of stay and patient care costs. *Academic Medicine* 1994, **69**, 489–95.

26 Urquhart, C. J. & Hepworth, J. B. Comparing and using assessments of the value of information to clinical decision-making. *Bulletin of the Medical Library Association* 1996, **84**, 482–9.

27 Wood, F., Wright, P. & Wilson, T. *The Impact of Information Use on Decision Making by General Medical Practitioners.* British Library Research and Development Report. Boston Spa: British Library, 1995.

28 Wood, F., Palmer, J., Ellis, D., Simpson, S. & Bacigalupo, R. Information in primary health care. *Health Libraries Review* 1995, **12**, 295–308.