Canadian Agency for
Drugs and Technologies
in Health

Agence canadienne
des médicaments et des
technologies de la santé

HTA

PRESS: Peer Review of Electronic Search Strategies

January 2008

Supporting Informed Decisions

Until April 2006, the Canadian Agency for Drugs and Technologies in Health (CADTH) was known as the Canadian Coordinating Office for Health Technology Assessment (CCOHTA).

Publications can be requested from:

CADTH
600-865 Carling Avenue
Ottawa ON Canada K1S 5S8
Tel. (613) 226-2553
Fax. (613) 226-5392
Email: pubs@cadth.ca

or download from CADTH's web site:
http://www.cadth.ca

Cite as: Sampson M, McGowan J, Lefebvre C, Moher D, Grimshaw J. *PRESS: Peer Review of Electronic Search Strategies*. Ottawa: Canadian Agency for Drugs and Technologies in Health; 2008.

Production of this report is made possible by financial contributions from Health Canada and the governments of Alberta, British Columbia, Manitoba, New Brunswick, Newfoundland and Labrador, Northwest Territories, Nova Scotia, Nunavut, Ontario, Prince Edward Island, Saskatchewan, and Yukon. The Canadian Agency for Drugs and Technologies in Health takes sole responsibility for the final form and content of this report. The views expressed herein do not necessarily represent the views of Health Canada or any provincial or territorial government.

CADTH is funded by Canadian federal, provincial, and territorial governments.

Legal Deposit – 2008
National Library of Canada
ISBN: 978-1-897465-38-7 (print)
ISBN: 978-1-897465-39-4 (online)
H0477 – January 2008

PUBLICATIONS MAIL AGREEMENT NO. 40026386
RETURN UNDELIVERABLE CANADIAN ADDRESSES TO
CANADIAN AGENCY FOR DRUGS AND TECHNOLOGIES IN HEALTH
600-865 CARLING AVENUE
OTTAWA ON K1S 5S8

Canadian Agency for Drugs and Technologies in Health

PRESS: Peer Review of Electronic Search Strategies

Margaret Sampson, MLIS PhD Candidate[1]
Jessie McGowan, MLIS PhD Candidate[2]
Carol Lefebvre, MSc HonFCLIP[3]
David Moher, PhD[1]
Jeremy Grimshaw, MBChB PhD FRCGP[2]

January 2008

[1] Children's Hospital of Eastern Ontario Research Institute, Ottawa ON
[2] Institute of Population Health, University of Ottawa ON
[3] UK Cochrane Centre, Oxford, UK

Reviewers

These individuals kindly provided comments on this report.

External Reviewers

Ellen T. Crumley, MLIS, AMIP, PhD (Provisional Candidate)
President
Health Info & Searching Practice Inc.
Edmonton, Alberta

Lindsay Glynn, BA, MLIS
Editor-in-Chief, Evidence Based Library and Information Practice
Public Services Librarian and Instruction Coordinator
Health Sciences Library
Memorial University of Newfoundland
St. John’s, Newfoundland & Labrador

K. Ann McKibbon, MLS, PhD
Associate Professor
McMaster University
Hamilton, Ontario

Peer Reviewer

Monika Mierzwinski-Urban, BA, MLIS
Information Specialist
CADTH
Ottawa, Ontario

CADTH takes sole responsibility for the final form and content of this report. The statements and conclusions in this report are those of CADTH and not of its panel members or reviewers.

Authorship

Margaret Sampson co-authored the funding application, supervised the project research assistant, co-developed the search strategy, participated in screening, extracted data, drafted the technical report, and participated in revisions.

Jessie McGowan co-authored the funding application, co-developed the search strategy, participated in screening, developed the survey methodology, supervised the development of the peer review forum and pilot, led or co-led presentations and feedback sessions with the community of systematic review librarians, and participated in the drafting and revision of the technical report.

Carol Lefebvre co-authored the funding application, advised on information science aspects of the project, facilitated feedback sessions with the community of systematic review librarians, and participated in revision of the technical report.

David Moher co-authored the funding application, supervised study personnel, advised on reporting standards and peer review, advised on methodology, and participated in revision of the technical report.

Jeremy Grimshaw co-authored the funding application, supervised study personnel, advised on guideline development and methodology, and participated in revision of the technical report.

Acknowledgements

We acknowledge the contributions of those who helped in the creation of this report. Janet Joyce, librarian and past director of Information Services at CADTH, served as an advisor on this project.

Elise Cogo, research associate, performed outstanding work on all aspects of the project. Isabelle Gaboury and Nick Barrowman advised on the statistical analysis. Raymond Daniel, assistant information specialist, provided technical support. Tamara Rader, librarian, provided writing assistance. Alex Korablin and David Mackie developed the Peer Review Forum web space. Special thanks to all the search experts who helped pilot or participated anonymously in the survey.

Conflicts of Interest

The authors have no conflicts of interest to report.

EXECUTIVE SUMMARY

The Issue

The quality of health technology assessment (HTA) reports depends on many factors. One of these factors is the evidence base from which the HTA is derived. The evidence base is created by gathering information from many sources and performing literature searches. Performing a high quality search of information resources will ensure the accuracy and completeness of the evidence base used in HTA reports. Currently, no review exists to tell us what elements of the search process have the most impact on the overall quality of the resulting evidence base.

Objectives

The objectives of the assessment are:

- to identify the elements associated with the accuracy and completeness of the evidence base found using electronic search strategies in different topic areas and apply this knowledge to HTA reports
- to determine the impact of errors in the elements of the electronic search strategy on the resulting evidence base
- to propose enhancements in the methods used for creating and evaluating search strategies to directly and positively affect the applicability of HTA reports.

With the goal of developing and validating a process of peer review for electronic search strategies, we considered tools that were developed in other areas of information retrieval that might serve as a basis for peer reviewing search strategies.

Methods

A systematic review, web-based survey, and peer review forums were performed. The systematic review was conducted to identify evidence related to quality issues and errors in complex electronic search strategies. Evidence was considered from any context, not only from research in systematic reviews and HTA searching.

The databases searched included Library & Information Science Abstracts (LISA, CSA interface) 1969 to May 2005; Cochrane Methodology Register & Cochrane Methodology Reviews (completed reviews only, The Cochrane Library 2005, Issue 2, Wiley interface); MEDLINE (OVID interface) 1966 to June week 1, 2005; PsycINFO (OVID interface) 1806 to June week 2, 2005; Cumulative Index to the Nursing and Allied Health Literature (CINAHL), (OVID interface) 1982 to June week 2 2005; HealthSTAR (OVID interface) 1987 to May 2005; and Health and Psychosocial Instruments (HAPI) (OVID interface) 1985 to March 2005. Efforts were also made to identify grey literature.

Because of the anticipated paucity of research evidence in some aspects of the electronic search, a web-based survey of expert searchers in systematic reviews and library and information studies was undertaken. The aim of the survey was to gather experts' opinions regarding the impact of search elements on the search results and the importance of each element in the peer review of electronic search strategies. The survey was conducted after the systematic review was completed, so that elements that were identified as potentially important in the review could be addressed in the survey. The original 14 elements studied in the review and five additional elements that were identified

during the review were included in the survey. After this, two peer review forums were held to discuss the results of the systematic review and survey.

Findings

A systematic review identified evidence on the importance of 14 of the 19 elements of the electronic search strategy that were initially considered. No evidence was found for two elements and from the three remaining elements, one additional element emerged as a result of the review. Although 26 tools were identified that could be used as checklists, none were validated for assessing electronic search strategies. Ten of these tools look at the conduct or reporting of the entire search (not just the electronic component).

Opinions were sought through a web-based survey for the elements that were considered in the systematic review. Fifty-eight respondents completed the survey. The elements were ranked into three tiers of importance based on an assessment of the potential impact of the elements on recall and precision. Elements that were rated as unimportant in peer review were dropped from further consideration. Based on the evidence of our findings from the systematic review, survey, and peer review forums, a process for validating the search strategy using a checklist and a peer review process was developed.

Conclusions

This work fills a gap in the assurance of the methodological quality of systematic reviews by contributing an evidence-based scale for the peer review of the electronic search strategy. The project has received support and participation from the information science community, and this approach to the peer review of search strategies has been supported by the Cochrane Collaboration's Information Retrieval Methods Group. A validated process - both transparent and robust - for peer-reviewing search strategies will improve the retrieval of the relevant information that forms the evidence base.

TABLE OF CONTENTS

APPENDICES – available from CADTH's web site www.cadth.ca

ACRONYMS AND ABBREVIATIONS

CINAHL	Cumulative Index to the Nursing and Allied Health Literature
HAPI	Health and Psychosocial Instruments
HTA	health technology assessment
LISA	Library and Information Science Abstracts
MeSH	Medical Subject Headings
PICO	population, intervention or exposure, comparison, and outcome
PRESS	Peer Review of Electronic Search Strategies
RCT	randomized controlled trial

GLOSSARY

Adjacency operators: See *proximity operators.*

Boolean logic: System of logical operators to join sets. Standard Boolean operators used in searching are "AND," "OR," and "NOT." Proximity operators imply "AND" and are another form of logical operator. Named after George Boole, a self-educated English mathematician.

Checklist: In this report, term used synonymously for other forms of evaluation such as questionnaires, scales, tools, instruments.

Check tag: Term routinely considered for use in indexing. In MEDLINE, usually included in Medical Subject Headings (MeSH) term field (but this is up to the vendor).

Controlled vocabulary: Consistent collection of terms chosen for specific purposes with explicit, logical constraints on intended meanings and relationships in a database.

Descriptors: See *subject headings.*

Double publication: Two or more published papers representing same research data or result, sometimes called duplicate publications. Double publication may be overt, with the other publications acknowledged or referenced, or covert, with acknowledgement sometimes deliberately disguised.

Duplicate: A redundant record pointing to the same full-text article. Records are usually not identical, because they may come from different databases and may differ in the treatment of authors' names or journal titles, indexing, and special fields.

EHTAS (Evaluating Health Technology Assessment Searches): Original project name. Name of project changed to Peer Review Electronic Search Strategy (PRESS) early in 2006 after consultation with local and international advisors.

Explode: Subject headings are arranged hierarchically in many thesauri. To explode a subject heading involves including a selected subject heading and all of the narrower terms that are below it in the hierarchy.

Filter: Search parameters designed to limit subject areas to a particular concept (focus of peer review should be to determine if use of filter is warranted, given the question.)

Fields: Searchable items in database, for example, authors' names, institutions, controlled vocabulary, titles, or abstracts.

Floating subheadings: Floating subheadings look for any subject heading that uses that subheading irrespective of which subject heading it is assigned to.

Free text: Normally words, phrases, or terms sought in title, abstract, or full text of document but this varies by database and vendor. See also *natural language*.

IQR: Inter-quartile range (25th to 75th percentile).

Irrelevant: In this report, “irrelevant” means “not meeting inclusion criteria of systematic review or HTA for which search is developed.”

Limit: System-based addition to search that is designed to exclude certain material not relevant to review. Examples include publication date, document type, and age group. (Focus of peer review should be to determine if use of limit is warranted, given the question.)

Listserv: Electronic mailing list where messages are distributed to all who subscribe to list. Most are based on a topic of mutual interest to subscribers.

Major heading: Subject heading that is designated as representing a main subject of document being indexed. In some interfaces, intention to retrieve only records where term is assigned as major heading is indicated by putting an asterisk in front of term – sometimes known as “starring.” Other interfaces may use terminology such as “restrict to focus.”

Modal rating: Most common rating.

Natural language term: Words, phrases, or terms sought in title, abstract, or full text of document. See also *free text*.

Negative impact on precision: Search error reduces ratio of accurate search results to inaccurate search results that are retrieved.

Negative impact on recall: Search error reduces number of relevant results that are retrieved.

Null retrieval: Search retrieval set with no records.

Peer review: Process of subjecting research work to independent scrutiny of qualified experts (peers); may be evaluated against certain standards (such as authorship guidelines).

Positive predictive value: Epidemiological term that usually refers to accuracy of diagnostic test. Computational equivalent of precision.

Precision: Proportion of retrieved items that is relevant. Equivalent to positive predictive value.

Proximity operators: Logical operators that specify the connected elements must not only both be present but must also be within specified proximity. Exact operators and their functions vary by system and include “NEAR,” “WITH,” “SAME,” and “ADJ.” Also called adjacency operators.

Recall: Proportion of relevant items in database retrieved by search. Usually known only in experimental situations although it can be estimated by statistical methods such as capture-mark-recapture. Also called sensitivity. Most searches for systematic reviews and HTA try to achieve highest practical recall, often at expense of precision.

Redundant: Search element that retrieves no additional records.

Relevant: In this report, defined as meeting inclusion criteria of systematic review or HTA for which search is developed.

Retrieval set: Records retrieved by search statement.

Search: In this report, "search" is electronic search strategy designed for retrieval from bibliographic or abstracting and indexing databases. Other elements of search plan for systematic review or HTA are database selection and selection of additional sources such as registries, hand or electronic searches of full text of journals and conference proceedings, communications with authors and manufacturers, web searching, electronic or manual checking of cited references.

Search performance: In this report, a measure of recall, precision, specificity, cost, or time.

Search query: See *search statement.*

Search statement: One line in electronic search strategy.

Search result: Anticipated or actual outcome of search term, statement, or strategy.

Specificity: Epidemiological term referring to accuracy of diagnostic test at correctly classifying negative cases as negative. Sometimes reported in assessments of accuracy of search strategies but it is not equivalent of precision.

Strength of research evidence: Validity of research underpinning any statement. If research evidence is strong, we can assume that underlying research is valid and based on appropriate research design.

Subheadings: Terms (sometimes called qualifiers) used with Medical Subject Headings (MeSH). See also *floating subheadings.*

Subject headings: Terms that make up the controlled vocabulary of bibliographic database. In MEDLINE, these are called Medical Subject Headings (MeSH).

Subject search: That part of search developed by searcher to address question (includes information about review topic). Filters and limits not developed by searcher may be added to this. Subject search should be main focus of peer review.

1 INTRODUCTION

1.1 Background and Setting in Canada

One of the many factors determining the quality of health technology assessment (HTA) reports is the evidence on which they are based. Performing a high quality electronic search of information resources ensures the accuracy and completeness of the evidence base used in HTA reports. Understanding which elements of the electronic search process have the most impact on the overall quality of the resulting evidence base will improve the HTA's accuracy.

This project will focus on the identification of issues associated with the accuracy and completeness of the evidence base used in HTA reports.

1.2 Overview of Technology

Checklists for validating aspects of the systematic review process have been developed, and some of these address aspects of the overall search plan.[1] What is lacking is a validated process for evaluating the quality and completeness of the electronic search strategy. The absence of such a process paired with a demonstrable level of error in reported searches[2] leaves this type of research open to debate over the quality of evidence on which the review is based. Without assurance of a bias-free and complete evidence base, the true outcomes of a systematic review cannot be ascertained.

2 ISSUE

With the goal of developing and validating a process of peer review for electronic search strategies, we considered existing checklists developed in other areas of information retrieval that might serve as a basis for peer reviewing HTA and systematic review search strategies. Anticipating the absence of adequate instruments, we also sought evidence regarding elements of the electronic search that could have a positive or negative impact on the performance metrics of recall and precision. Research evidence on the impact of search errors on search performance would be the most compelling. We also sought research evidence through a systematic review of the literature. We expected some gaps in the evidence and so we considered reports from the literature on the prevalence of errors of each type and the theoretical explanations of these impacts. Finally, we used a survey to solicit experts' opinions from the community of systematic review and HTA searchers and from other expert searchers in librarianship and information studies.

The electronic search is the focus of this review and of the Peer Review of Electronic Search Strategies (PRESS) checklist, but it is only one aspect of a comprehensive search for systematic reviews and HTAs. Other aspects include hand-searching, searching reference lists, Internet searching, grey literature searching, and contacting authors. These are essential components of the systematic review literature search. We thought that it would be difficult to peer review these other methods without a standard procedure and chose to look exclusively at the electronic search for this project.

3 OBJECTIVES

The following are the objectives of the assessment:

- to identify elements associated with the accuracy and completeness of the evidence base found by electronic search strategies used in HTA reports
- to determine the impact of errors in any of these elements in the electronic search strategy on the resulting evidence base
- to propose enhancements in the methods used for creating and evaluating the electronic search strategies used in reviews to directly and positively affect the applicability of HTA reports.

4 EVIDENCE REVIEW

4.1 Methods

A systematic review was conducted for evidence related to two questions.

- Are there any existing checklists that evaluate or validate the quality of literature searches in any discipline?
- What are the elements that relate to quality or errors in search strategies? These articles need to have performance indicators or measures (such as recall or relevance).

The research plan presented in the grant application served as the study protocol.

4.1.1 Literature search strategy

The electronic search strategy was developed initially in the bibliographic database Library and Information Science Abstracts (LISA). One researcher (EC) developed the search, and two co-investigators reviewed and revised it (MS, JM). The search strategy was adapted for each of the other databases searched. Material published from 1980 and onwards was sought, reflecting the rise in widespread use of electronic searching. Languages were limited to those understood by the review team: English, French, Italian, and Spanish.

The databases searched were LISA (CSA interface) 1969 to May 2005; Cochrane Methodology Register & Cochrane Methodology Reviews completed reviews only (The Cochrane Library 2005, Issue 2, Wiley interface); MEDLINE (OVID interface) 1966 to June week 1, 2005; PsycINFO (OVID interface) 1806 to June week 2 2005; Cumulative Index to the Nursing and Allied Health Literature (CINAHL) (OVID interface) 1982 to June week 2 2005; HealthSTAR (OVID interface) 1987 to May 2005; Health and Psychosocial Instruments (HAPI) OVID interface) 1985 to March 2005 (Appendix A).

Grey literature was identified through correspondence with information specialists and other experts; by searching The Cochrane Methodology Register, which contains conference abstracts; and by searching our personal databases of information science research accrued over the years, including unpublished material such as presentations, dissertations, and pre-publication manuscripts.

The references were imported into a Reference Manager database, and duplicate records were removed. The remaining records were uploaded to SRS^{TM}, a web platform for systematic reviews.

After screening and data extraction of this material, the reference lists of included studies were checked to capture literature on those elements for which we had found fewer than five studies. These elements were proximity operators, the Boolean operator "NOT," conceptualization, organization, irrelevant terms, truncation errors, wrong line number specified, explosion without the existence of a narrower term, searching additional fields, redundancy without rationale, and combining index terms with free-text terms on the same line.

4.1.2 Selection criteria

Initially, the bibliographic records (title, abstract, and indexing terms) were assessed as relevant or not relevant to peer review of electronic search strategies, and reports of primary research or secondary reports (review articles, tutorials) citing supporting evidence (such as recall or precision) were selected. Articles identified using these broad criteria were then examined, and articles that presented an evaluation checklist for search strategies, or presented primary evidence on the impact of searching techniques on search results, or presented a theoretical discussion on the impact of searching techniques on search results were selected for inclusion in the systematic review.

4.1.3 Analytical framework

Information retrieval is at the forefront of library science research, yet much of the research is descriptive and not all aspects of practice have a firm evidential basis.[3] Thus, theoretical work and experts' opinions were considered in the analytic framework.

An initial assessment of the literature was based on a pragmatically derived list of purported search errors (elements) previously examined by the authors.[2] Additional types of errors or practices that have a negative impact on search performance were identified during the systematic review of the literature. The original elements and those added during the systematic review were included in a survey of interested professionals.

Evidence from all sources was summarized and classified into one of three tiers based on the balance of research evidence, theory, and experts' opinions.

Elements were retained when there was evidence that peer review, through the detection of specific errors, could improve the performance of the electronic search and potentially improve the evidence base for a systematic review or HTA or improve the efficiency of its development. Where possible (i.e., when the factors to be assessed in the peer review process were similar), elements were combined.

Lastly, for each final element, a summary of the evidence and the focus for evidence-based peer review was prepared. Elements are presented in order by tier.

4.1.4 Selection method

For the assessment of eligibility, the training of reviewers and the calibration exercises were done in three stages until the cumulative kappa score was within the acceptable range of agreement according to Landis *et al.*'s criteria.[4] After these three stages were completed, the titles and abstracts were screened by one of the three reviewers for potential eligibility, but a second reviewer was needed to confirm ineligibility before a record was excluded. Articles appearing to be potentially relevant were retrieved, and two reviewers assessed each of the full reports, arriving at a consensus on eligibility.

4.1.5 Data abstraction strategy

Another calibration exercise was performed for data extraction. Three articles that seemed to address a few of the elements were chosen, and all three reviewers extracted data from them. The results were compared, and very good agreement was found among the three reviewers. One reviewer did all subsequent data extraction, and a second reviewer verified it.

For each included study, it was determined which of 14 elements of the electronic search were addressed. Those elements were spelling mistakes, missed spelling variants, errors in truncation, problems in the logic or organization of the search, misapplication of the Boolean logical operators (“AND,” “OR,” “NOT”), wrong line number, subject headings and natural language terms combined in a single search statement, subject headings missing, natural language terms missing, irrelevant subject headings or explosion of subject headings to include irrelevant narrower terms, inclusion of irrelevant natural language terms, explosion of subject headings exploded even though no narrower terms exist, and redundancies in the search strategy. We also sought evidence on the impact of failing to adapt the search strategy to each database. A final open-ended question allowed for additional comments to be noted.

4.1.6 Data analysis methods

Data were summarized descriptively and synthesized qualitatively. Evidence came from three sources. The first source was the research on the impact of a search error on search parameters (such as recall and precision). The second source was the theoretical discussion of impact on search parameters. Finally, evidence regarding the prevalence of the search error was considered.

Research and theoretical works that considered the impact of errors on search parameters were classified according to the main parameters affected by the error. The primary parameters considered were recall and precision, which are important parameters in assessing information retrieval, and are analogous to sensitivity and positive predictive value. While other parameters may be of greater interest in other aspects of information retrieval, these two relevance-based parameters are most germane to the information retrieval task of systematic review searching where a formal dichotomous determination of relevance is made.[5] Recall determines the integrity of the evidence base. Precision determines the time and cost of screening the search results to extract the evidence base, but it also negatively affects the likelihood of correctly identifying relevant records from the retrieved record set. The secondary parameters were time, cost, and specificity, which are mostly associated with precision.

The research evidence was sparse for many search elements, and the research designs varied. For these reasons, neither quality assessment nor quantitative synthesis was undertaken.

Thus, the analytic approach was a dichotomous assessment of the presence or absence of research or theoretical evidence. This, paired with experts’ opinions from survey data, was used to classify errors into three tiers. The first tier consisted of search elements that the majority of experts thought had an impact on recall and thus were important in systematic review searching. The second tier consisted of search elements with mixed support or where precision was the main parameter affected. The third tier consisted of those elements largely unsupported by the literature review or by experts’ opinions.

For each element, which, on the balance of the evidence, was considered to be important for peer review, the salient features that would provide guidance to peer review were extracted and summarized. For instance, if subject headings were considered to be a first tier element, those aspects of usage that would improve recall were summarized. These could include appropriate methods to identify subject headings or best use of thesaurus features like exploding or applying subheadings that could help or hinder optimal retrieval.

4.2 Results

4.2.1 Quantity of research available

In all, 9,155 records were identified for screening, of which 256 full-text articles were obtained for more assessment. One hundred and thirteen articles were eligible for some aspect of the systematic review (Figure 1). A list of excluded articles can be found in Appendix E.

Figure 1: Selected reports

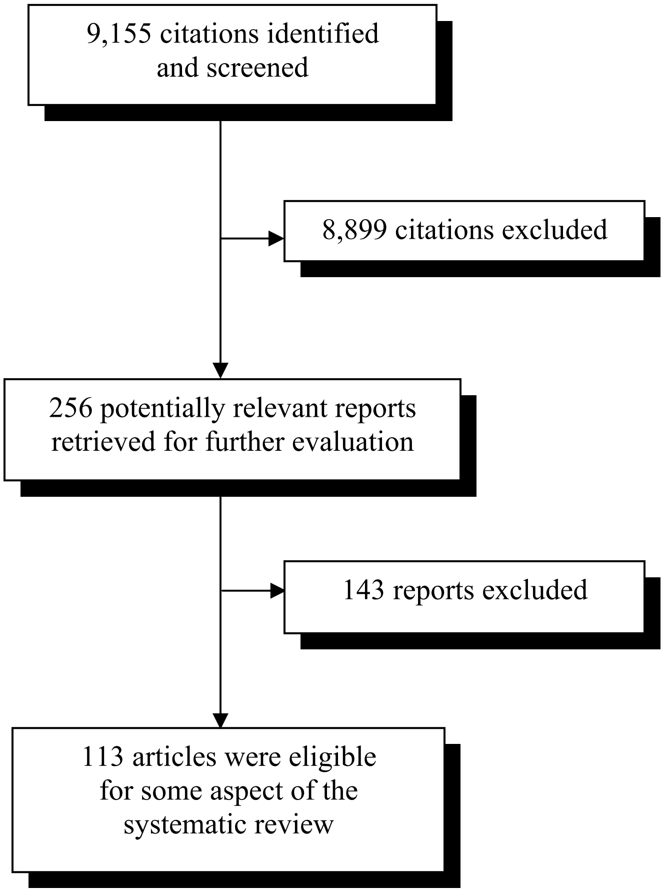
9,155 citations identified and screened
8,899 citations excluded
256 potentially relevant reports retrieved for further evaluation
143 reports excluded
113 articles were eligible for some aspect of the systematic review

Table 1 presents a summary of the amount and type of evidence identified from the literature review for each element. Elements are arranged into tiers, according to their final classification. The amount of evidence varied, with more than 120 papers addressing the consequences of missed search terms, and little or no identifiable evidence on aspects such as redundancy without rationale, subject heading exploded when no narrower terms exist, or index and free text combined on a line.

Table 1: Evidence for elements identified from systematic review of literature

First Tier Elements					
Element	**n**	**Evidence type (n)**			**Main impact**
		R	**T**	**F**	
Search not adapted for each database	27	17	14	1	recall
Conceptualization	23	10	12	4	recall
Logical operator errors	40	22	19	10	recall, precision
Boolean	22	9	14	3	recall, precision
Proximity operators	8	2	4	1	recall, precision
NOT	6	3	3	0	recall, precision
Missed index terms	81	56	31	7	recall, precision
Spelling errors	13	11	7	9	recall, precision
Second Tier Elements					
Missed free text terms	71	51	29	10	recall, precision
Limits used inappropriately or missed	31	13	15	2	recall, precision
Irrelevant index terms	3	2	1	1	precision
Truncation errors	2	1	0	1	precision
Irrelevant free text terms	4	0	2	1	precision
Third Tier Elements					
Failure to use additional fields	13	6	6	1	recall, precision
Organization of search	6	1	6	0	recall, time-cost
Redundancy without rationale	2	0	2	0	recall
Subject heading exploded when no narrower terms	0	0	0	0	none
Index and free text combined on a line	0	0	0	0	none

Type of evidence: research (R), theory (T), and frequency of error (F).

The main impacts that were examined were recall (or sensitivity), precision (or positive predictive value), specificity, cost or time, and any discussion of importance in peer reviewing (Appendix B).

The types of evidence considered were research evidence regarding search performance, theoretical rationale for impact on search performance, and frequency of error in a particular setting, such as catalogue searches by undergraduate students or searches by medical residents to answer clinical queries.

5 SURVEY

5.1 Objectives

We planned to use a web-based survey of experienced searchers in systematic reviews and library and information studies, because of the anticipated paucity of research evidence in some elements of the electronic search. The aim of the survey was to gather experts' opinions regarding the impact of search elements on the search results and the importance of each element in the peer review of electronic search strategies. The Ottawa Hospital Research Ethics Board gave approval. The survey was conducted after the systematic review was completed, so that elements identified as potentially important during the review could be addressed in the survey. The 14 original elements studied in the review and four additional elements that were identified during the review were included in the survey.

5.2 Methods

5.2.1 Survey development

Each element was presented in question form, as it might be presented in a peer review assessment – for example, "Are any spelling variants missing?" (Appendix C). The question was followed with a definition and in some cases, an example. The survey respondents were asked to rate on a four-point scale (ranging from nil to large) the potential negative impact on recall; to rate on a five-point scale (ranging from strongly disagree to strongly agree) whether problems with the use of the element might indicate the searcher's unfamiliarity with aspects of searching, and to rate on a five-point scale the importance of considering that element in the peer review of a search strategy. The respondents were asked to provide their impression of the level of evidence on the importance of the element. This was based on seven response categories; four relating to the strength of research support, while the other three related to support by experts' opinions, a self-evident impact, or absence of support. A final question allowed respondents to nominate any additional errors that should be considered, with the respondents' assessment of the impact and impression of level of evidence and importance in peer review.

The survey was pilot tested on 10 participants and then refined based on feedback from the pilot testing (Appendix C).

5.2.2 Sampling frame

Survey respondents were recruited from eight listservs dealing with systematic reviews and medical librarianship (Table 2).

The recruitment message described the eligibility criteria for survey participants, expected time to complete the survey, its purpose, and funding source. If respondents requested more information, a version of the survey invitation containing a summary of the PRESS project was sent. Those with detailed questions were sent individual email responses, in addition to the survey invitation.

Table 2: Listservs used in survey recruitment

Listserv	Audience
HTAi SPIG-IR	Health Technology Assessment international Special Interest Group for Information Resources (SPIG-IR)
InterTASC ISSG	InterTASC Information Specialists' Sub-Group (ISSG), UK-based HTA information specialists
MEDLIB-L	An Email List for Medical Librarians
CHLA	Canadian Health Libraries Association / Association des bibliothèques de la santé du Canada
UNYOC	The Upstate New York and Ontario Chapter, Medical Library Association
OVHLA	The Ottawa Valley Health Libraries Association / L'Association des bibliothèques de la santé de la vallée de l'Outaouais
Cochrane IRMG	The Cochrane Information Retrieval Methods Group
Cochrane TSC	The Cochrane Collaboration's Trials Search Coordinators list

The demographic characteristics that were recorded were years of searching experience, years of systematic review experience, estimated number of systematic reviews or HTA reports undertaken, formal training and degrees, and country of residence (Appendix C). Although the letter of invitation described eligibility criteria, no respondent was excluded based on the demographic characteristics that they reported.

5.2.3 Survey administration

The survey was conducted using *SurveyMonkey,* a web-based software application. The survey took place between September 1 and September 19, 2005. Respondents to the listserv recruitment letters were sent an invitation containing a tracked link to the survey via email. The survey invitations included an option to be removed from the list. The consent form was the first page of the survey (Appendix C). Survey responses were filtered based on consent to participate. Reminders were sent after one week via the *SurveyMonkey* list manager, to those who had expressed interest, but who had neither completed the survey nor withdrawn. Those who consented but completed only part of the survey were sent weekly reminders. Late in the survey we learned that some institutional firewalls were blocking emails from the *SurveyMonkey* list manager. The investigator's institutional email account was then used to send a generic link to the survey and a reminder to any remaining non-responders or partial responders.

5.2.4 Data cleaning strategy

Upon completion, survey results from consenting respondents were exported from *SurveyMonkey* to a spreadsheet. The data were re-coded, and duplicates were removed.

Only surveys where the respondents consented to participation were examined (n=70). Surveys with at least one response beyond the demographic data were deemed to be valid. There were 58 valid surveys. Nine submissions with demographic data only were excluded from the analysis. One respondent submitted two such responses, and two people completed one valid survey and one with demographic information only.

The variable "Country" was added after the survey was launched. For the first five cases, data on country were inferred from the email address or telephone country code provided by the respondent.

Text responses for years of searching experience and number of systematic reviews or HTAs in which the respondent had participated were cleaned. Where the respondent specified a range, the mid point was entered. Where the respondent indicated a minimum (e.g. 20+ years), the minimum was entered. Where the respondent entered an approximate number (e.g. about 12), we entered that number. Academic preparation (or level of education achieved) was recorded in the following categories: library or information science training (all degrees), health care practitioner training (including MD medical doctor, ND naturopathic doctor, and nursing), epidemiology training (MSc or PhD), or other. When respondents indicated multiple degrees, we recorded their first response.

5.2.5 Data analysis methods

Except for the element addressing the impression of the level of evidence on the importance of a specific element (item d for each survey element), all elements were ranked on a dimension according to median score. For elements with the same median, those with the smaller inter-quartile range (IQR, indicating greater consensus) were ranked higher. Where there was still a tie, the elements were considered equal. The research evidence dimension was re-coded so the answer “No support” was scored as 0, “No research evidence but self-evident” was scored as 1, and all else was scored as 2, indicating some level of research support. The ranking was then made based on the percent of respondents who thought that there was some research support for the element.

5.3 Expert Survey: Results

Fifty-eight respondents (expert searchers) completed the survey with usable responses. Because the number of individuals reached by the listserv announcements is unknown, the participation rate cannot be determined.

The mean number of years of searching experience for the respondents was 12.6. The mean number of years of experience with systematic reviews or HTA was 5.3. Respondents reported that the number of years of involvement in systematic reviews or HTA projects was 18.4. The percent of respondents with a library science degree or similar was 76.9. Nearly 85% of respondents were from Canada, the UK, or the US (Table 3).

5.3.1 Survey rankings of elements

The survey ratings of electronic literature search errors clustered into three tiers (Table 4). The first tier had only first and second place rankings on the three variables to which the survey respondents gave the most weight: impact on recall, impact on precision, and importance in peer review. These elements were perceived to have research support or be self-evident.

The middle tier had strong support. The examination of the actual scores shows these to be high, even though lower or less consistently high than for the first tier elements. The survey respondents supported inclusion of these elements in peer review.

The elements in the final tier had no first or second place ranking on recall, precision, or importance. Respondents perceived them as neither supported by research evidence, nor a marker of inexperience in searching.

Table 3: Characteristics of survey respondents

Year of searching experience		
Mean		12.56
Standard deviation		7.94
Number responding		58
Years of experience with systematic reviews or HTA		
Mean		5.30
Standard deviation		3.77
Number responding		58
Estimated number of systematic review or HTA projects		
Mean		18.38
Standard deviation		23.56
Number responding		58
Academic preparation or formal training	**Number**	**%**
Library science degree or similar	50	76.9
Epidemiology	3	4.6
Health care practitioner (physician, nurse)	2	3.1
Other	4	6.2
Not reported or not interpretable	6	9.2
Country of residence	**Number**	**%**
Canada	27	47.5
UK	15	25.4
US	7	11.9
Denmark	2	3.4
Spain	1	1.7
Norway	1	1.7
Switzerland	1	1.7
South Africa	1	1.7
New Zealand	1	1.7
Malaysia	1	1.7
China	1	1.7

Other factors reported by respondents included whether an error in an element indicated a lack of familiarity with the search and whether the respondent thought that there was research support for the element's importance. These were not considered in making the classification by tier, but are presented in Table 4.

6 DATA ANALYSES AND SYNTHESIS

Evidence from the systematic review and survey assessment is presented for each element considered (Table 1 and Table 4).

Table 4: Search element survey ranking results

Element	Important in peer review	Recall	Precision	Unfamiliarity	Research evidence (impressions)	
Adapted for each database	1	1	1	1	6	*
Boolean errors	1	1	1	1	12	*
Line number errors	1	1	1	5	18	*
Translated for each database	1	2	2	5	1	*
Subject heading missing	2	1	2	1	2	*
Spelling errors	2	1	2	7	15	*
Free text missing	2	2	3	3	3	
Irrelevant limits	2	2	4	3	8	
Spelling variants missing	2	2	5	2	4	
Irrelevant subject headings	2	4	1	2	9	
Truncation errors	3	2	2	3	14	
Limits missing	3	3	2	3	5	
Irrelevant free text terms	3	6	2	3	13	
Redundancies	3	8	7	5	17	†
Organization of search	4	3	6	7	10	†
Additional fields	4	5	5	5	7	†
Subject headings and free text combined	4	6	7	6	11	†
Irrelevant explosions	5	7	8	4	16	†

*These 6 elements had only first or second place ranking on recall, precision, or importance, and are considered to be first tier elements. †These 5 elements had no first or second place ranking on recall, precision, or importance, and are considered to be third tier elements.

6.1 First Tier Elements

6.1.1 Conceptualization

At issue is whether the electronic search strategy includes the most important elements of the clinical question, with neither too few nor too many concepts introduced. This element focuses on the concepts, not how well they are expressed in search statements.

Of 23 papers related to this element, 17 addressed impacts on recall, and six considered precision. Ten research papers were identified.[6-15] Four papers reported on the frequency of errors.[7,16-18] Twelve papers discussed conceptualization but did not present research evidence related to this element.[16,18-28] Most of these focused on recall.

Almost all research and commentary supported the need for effective conceptualization. One study, however, noted the low overlap in terms selected by experienced searchers. This could be considered as mildly refuting evidence of the importance of conceptualization in any but a high recall situation.[23] Several writers in this sample dealt with the question negotiation process. For example, one describes questioning techniques to elicit the nature of the problem and cautions against accepting a list of key words offered by the client.[25] Jokic summarizes, “In other words, once the searcher understands the request well enough to answer it, a plan is developed for the search – a search strategy.”[16]

Fidel classified the actions made by professional searchers in developing a search as conceptual or operational.[8,9] A conceptual move modifies retrieval by changing the meaning of the concept that it represents. Conceptual moves were most often used to increase recall, while operational moves (generally invoking database features such as limiting to major subject headings, to documents of a certain form, to a particular language or publication date, restricting free text searches to certain fields, using adjacency operators instead of "OR") were generally used to improve precision. This was thought to indicate that searchers were reluctant to narrow the meaning of a request.

Although no research evidence is presented, over-specification (e.g., including too many concepts)[11,25] is listed as a common pitfall that reduces recall. Missing concepts (presumably to be connected with other concepts of interest) were identified as being the cause of retrieval failure for 25% of articles not retrieved in a full-text setting.[11] Precision was improved by using questions structured in the "population, intervention or exposure, comparison, and outcome" (PICO) format.[6] One reason for low precision was searching only one concept from the question.[12] Highly complex questions yielded result sets with higher precision than did those of low complexity (odds ratio = 2.16, P <0.05). Recall was lower, but not significantly lower (odds ratio = 0.82, P >0.05).[13] Several papers dealt with structuring the search statement in the PICO format.[7,19,28] Three papers advocated the use of diagrams to aid in the conceptualization of the search.[24-26]

Several research articles reported data on the prevalence of conceptual errors. Mularski and Bradigan found that before training, 47% of medical students made conceptual errors in identifying essential concepts in the statement of a search problem.[17] Bradley *et al.* assessed the frequency of errors in conceptualizing the search and found that between 60% and 100% of participants were making such an error, even after training, in situations where the question posed contained some ambiguity and searchers had difficulty determining if the question was about therapy or prognosis.[7] Novices were found to make errors by neglecting concepts and by adding unnecessary concepts. Jokic found that approximately half of searches were conceptually sound, regardless of the non-librarian searcher's basic ability to use the system independently. Those who were able to search the system independently were more likely to create searches that were conceptually sound, but expressed the concepts poorly (29% versus 17%). The independent searchers produced fewer conceptually flawed searches than did those who needed technical assistance to execute the searches (14% versus 0%).[16]

Two papers discussed the peer review of search conceptualization. The first advocated diagramming of the strategy as a useful supplement to the peer review discussion.[25,26] In the second, the Norris Library used the correct interpretation and use of stated information requirements as one of four elements of peer review. Glunz and Wakiji state that "despite a well-designed search request form and a thorough search interview, interpreting the request can still be difficult. The goal of the reviewer is to check whether the searcher has identified the components of the request and separated them in logical groups."[24]

Survey respondents rated the translation of the search question into search concepts as being important (Table 4).

6.1.2 Spelling errors

Thirteen papers addressed spelling errors.[7,12,15,16,25,29-36] Eleven provided research evidence[7,12,15,16,29-32,34-36] and seven were theoretical papers,[7,12,16,25,33-35] two of which did not also present research evidence. Of the research papers, most focused on factors accounting for searches that failed to retrieve relevant items – a recall problem. One research paper found a negative effect of spelling errors on precision.

In a study of web searching, the number of misspelled terms was also significant as a predictor of search performance.[31] Typographical errors lowered search performance – some found that such errors were usually noticed and corrected quickly.[30] A misspelling embedded in a long search statement, however, may go unnoticed if elements are connected with the "OR" operator. The problem would be more apparent in a string of terms connected with "AND."[25] Errors in system commands may retrieve some hits, but they may be unrelated to the intended query. An example of a fatal error in DIALOG, provided by Bellardo and Saracevic, is typing "S 15/ENG" rather than "S S15/ENG" —resulting in English-language papers having the number 15 in the record, rather than limiting set 15 to the English language only.[23]

Up to 6% of null retrievals in catalogue searching were due to spelling or typing errors.[29] Approximately one-quarter of end-user searchers made typing errors.[16] Walker *et al.* found that spelling errors in subject words accounted for only one of 172 unproductive searches (those retrieving no records) by clinicians searching MEDLINE.[15] Bysouth reported that typing or spelling errors occurred in 25% of basically flawed strategies conducted by research scientists.[36] Mitchell *et al.* found misspellings in 8% of failed no-hit searches by biochemistry students.[14] Sewell and Teitelbaum noted many corrected typing errors, but searchers seemed unable to correct all such errors.[37] Sewell and Bevan (as cited by Wildemuth and Moore) analyzed errors made by pharmacists and pathologists searching TOXLINE and MEDLINE. The most common errors were related to misspelled terms and misuse of the controlled vocabulary.[30]

Spelling errors can also occur in the bibliographic record. Ray and Vermeulen (as cited by Barroso *et al.*) found that most (71%) of the misspelled words that they identified in a MEDLINE study occurred in the abstract, and that most of these would not be retrieved by using the correctly spelled word unless the subject heading was also searched.[34]

6.1.3 Logical operator errors

Logical operators in search strategies are used to place the individual terms or concepts in mutual relationships.[16] Placing the terms into the wrong relationship would affect the search. Boolean logical operators ("AND," "OR," and "NOT") are considered, as are proximity and adjacency operators ("NEAR," "WITH," "SAME," "ADJ").

Of research papers reporting on logical operators, 17 dealt with recall, 13 with precision, and two with specificity.[10,11,15,16,18,20,23,30,31,36-49] Nineteen articles provided a theoretical rationale for the impact of the error, or provided search instruction or advice.[9,16-21,25,26,28,33,47,50-56] The frequency of the error in a particular population was reported in 10 papers.[15-18,30,36,37,43,57,58]

Research on logical operators focused on the use and misuse of Boolean logic and proximity operators. Bysouth examined searches by research scientists and noted the use of various features, errors in fundamentally flawed searches, and ways to improve poor strategies. Boolean operators were used in 78% of strategies: "AND" in 76%, "OR" in 43% (usage figures for the "NOT" operator were not reported).[36] Nested logic was used in 15% of searches. In flawed searches, the excessive use of "AND" was a problem in 11%, the use of mixed logic without proper nesting was a problem in 11%, and the illogical use of Boolean operators was seen in 4%. Research scientists used proximity operators in 54% of searches, and a further 11% of searches could have been improved by their use. This was the only source of missed opportunity noted among the logical operators.

Omissions (failure to use an operator where it could have aided the search) were more common than misused operators. The frequency of use of Boolean operators by third-year medical students showed that "AND" was commonly used to combine concepts, "OR" was less often used, and "NOT" was infrequently used.[12] Jokic found the same pattern of use in a population of moderately experienced end users at a university.[16] Proximity operators were rarely used. McKibbon noted that librarians and experienced searchers used the "OR" operator approximately three times as often as novice searchers.[46] Sewell and Teitelbaum note the infrequent use of "OR" and the absence of "NOT." They identified an implied "OR" where searchers would enter "PHENELZINE AND AGORAPHOBIA" followed by "PHENIPRAZINE AND AGORAPHOBIA," when this could have been represented by (PHENELZINE OR PHENIPRAZINE) AND AGORAPHOBIA. Errors in the use of "OR" occurred in eight of 352 sessions. No errors were seen in the use of "AND."[37] Kirby and Miller found the same pattern of usage of "AND" and "OR" by untrained end users, but the pattern of usage was similar between successful and unsuccessful searches.[10] One exception to this pattern of infrequent use of "OR" occurred in a study by King, who found that users of PaperChase tended to use "OR" nearly as often as "AND," regardless of the experience level.[58] This was attributed to the interface, which provided several mechanisms for performing an "OR" operation. Another study showed that users of a structured interface for connecting terms outperformed those using a simple interface who had received instruction in Boolean logic, and that such instruction was counter-productive when a structured interface was used.[44]

Of missed opportunities involving Boolean operators, "OR" was missed by 20% of the student searches assessed.[30] Illogical Boolean combinations were identified in the searches of 41% of student searches. Similar patterns of operator usage were seen in a study with web search engine users where "AND" was used most often, followed by "NOT," and "OR" was used rarely.[31] This study by Lucas and Topi, however, found "AND" to be missed more often than "OR." First order correlations between the number of incorrectly used operators and the average standardized relevance were not statistically significant.

In an analysis of transaction logs from an Internet search service, Jansen *et al.* studied the frequency of use of logic and modifiers and conducted a failure analysis to identify trends among users' mistakes.[57] "AND" was used in 8% of all queries, and 32% of these uses were erroneous. The "AND" operator was often used as a conjunction, as in "college and university harassment policy." "OR" and "NOT" were used in less than 1% of queries and were used incorrectly 26% and 37% of the time respectively. Walker *et al.* found that "ANDing" redundant or repetitive terms representing the same search concept accounted for 3% of unproductive searches by clinical end users.[15]

In Mitchell *et al.*'s study, medical students' logic errors appeared in 13% of the 208 "failed" searches, which retrieved no relevant hits (and 11% of the total 829 searches). Based on this, an emphasis was placed on instruction in the principles of Boolean logic.[12]

Improvement in the ability to use Boolean operators after training has been documented. In a training program for medical students, Mularski and Bradigan found that correct use rose from 44% on a pre-test to 91% on the post-test. The "OR" operator proved to be the most difficult with a high level of residual error after training.[17] Lucas and Topi found that web searchers could improve their search performance in a simple interface with even limited training in basic Boolean logic.[44] Training consisted of interactive web-based instruction in the Boolean operators, using quotation marks to construct phrases, and using parentheses to form query clauses. Improvement was seen in the ability to provide a correct answer to a search request, decreased time spent obtaining an answer, increased rating in ease of use of the search environment and increase in participants' confidence in their abilities. In assisted interfaces (i.e., where terms were entered in text boxes, and boxes were joined by picking operators from pull-down lists), receiving minimal training in Boolean logic was counter-productive.

In an error analysis involving 76 undergraduates, Nahl found that the two most common errors were the interrelated problems of non-probabilistic logic and semantic leakage.[18] Non-probabilistic logic ignores the formal requirements of Boolean logic by including all terms. Semantic leakage "occurs when searchers produce a search query that represents their ordinary understanding of the 'aboutness' of a topic and ignore the syntax of combining concepts in a search statement." Students were randomly assigned to look at one of two versions of written Boolean instruction. The group that received more intensive feedback had higher Boolean logic scores by 40% and was more confident in searching than those taught with basic instructions.

Proximity operators perform as a narrower form of "AND" because they require not only that all terms be present but also that they must be within the specified proximity. Kristensen studied a set of 26 searches performed in a full-text database under three conditions.[41] First, concepts were joined with "AND" (document co-occurrence). Second, proximity operators were used to restrict the occurrence of the concepts to the same paragraph (paragraph co-occurrence). The third and most restrictive form was sentence co-occurrence. With document co-occurrence representing the baseline measure for retrieval, paragraph co-occurrence reduced recall from 1.0 to 0.39, and sentence co-occurrence reduced it to 0.26. Precision was 0.51 with document co-occurrence, increasing to 0.67 with paragraph co-occurrence and 0.69 with sentence co-occurrence. Kristensen notes that the text used in the study (newspaper articles) tended to have short paragraphs, which may have led to lower recall than would occur in other contexts.

Tenopir and Shu searched with progressively more restricted proximity requirements and concluded that proximity operators are important in full-text searching and essential when the full-text environment has no controlled vocabulary. They noted a relationship between improved precision and lower recall when closer proximity was required.[49]

Keen explored the circumstances in which proximity is useful, the performance effect, and the frequency of use by searchers.[42] Keen followed Tenopir and Shu's method of progressively more restrictive searching with similar results.[43] Keen concluded that queries involving phrases and variations in wording often were best addressed with proximity searching, as were searches using frequently occurring terms that would otherwise result in large retrievals.

One challenge facing searchers is the range of terminology and functioning of operators across systems. McJunkin studied the effect of proximity operators in titles on recall and precision, and concluded that "although the results of this study seem to support the use of adjacency operators to improve searching effectiveness, a user for whom absolute recall is more important may wish to use a broader search strategy."[40] McJunkin also notes, however, that when text words are discipline-specific, proximity operators improved precision with little degradation in recall.

The Boolean operator "NOT" can reduce recall. Deacon *et al.* provide an example in which the "NOT" operator excluded burns from a child safety query and excluded a resource that was about a range of safety topics, including burns.[45] Jenuwine and Floyd also documented several retrieval failures attributable to using "NOT" to exclude children beyond a certain age.[39]

The use of Boolean operators qualified as a first tier element for peer review, based on respondents' assessments of its potential to negatively affect recall and precision and strong agreement of the element's importance in peer review.

In summary, Boolean logic can be mysterious to the novice searcher, and mistakes can be catastrophic in the search. Search logic is teachable, however, and amenable to peer review. Used effectively, Boolean operators can be used to construct complex searches.

6.1.4 Search strategy adapted for each database

Search strategies developed for one database need to be adapted to the indexing structure, limits, and special features available to another database. When the database searches are run on different search platforms, the variations in command syntax must also be considered.[26,59]

Although many papers addressed the need to tailor the search for databases,[9,20,22,24-28,34,36,42,43,49,51,53,57,59-69] there was virtually no research on the impact of failing to do so. One study undertook database searches in three databases using tailored strategies and analyzed the reasons for retrieval failure.[69] They did not report the retrieval implications of re-running the same search in all databases.

When searching more than one database, the adaptations needed are the subject headings,[20,34,59,61,63,65,69] including those that describe study methodology,[59,63,69] database features such as limits, and special fields, which typically vary by database but may also vary by platform (e.g., OVID versus Dialog). Proximity operators in particular vary in syntax and capabilities across systems.[34,42,43,49,51,66] Even translation from OVID's MEDLINE to PubMed requires re-development of the search because of differences in feature availability, even though PubMed includes all MEDLINE records.[60]

Survey results placed search adaptation in the first tier, and this element was identified as having the potential to make a large impact on recall and precision (Table 4). There was strong agreement that failure to adapt the search for different databases indicated a lack of familiarity with searching. Survey respondents thought that peer reviewing this element is important.

6.1.5 Wrong line number

Systematic review searches rely on Boolean searches. Including the wrong line number in a Boolean statement or limiting the wrong line, whether through typographical or logical error, could affect search performance. No research evidence on the impact or prevalence of using the wrong line number was found. Line number errors qualified as a first tier element for peer review, based on respondents' assessments of its potential to negatively affect recall and precision and strong agreement on the element's importance in peer review in responses to the survey (Table 4). The importance of this element, more than any other in the survey, was considered to be self-evident by respondents, and this is congruent with the paucity of research evidence — there are better questions to research.

6.1.6 Subject headings missing

Eighty-one papers discussed the importance of subject headings. Fifty-six involved research, and these are given the most emphasis here.[7,8,10-12,14-16,20,23,24,27,30,34,35,37-39,46,48,59,62,66,67,69-100] Thirty-one included theoretical discussion, [7,9,12,16,17,19-22,24-28,33-35,52-56,62,63,68,72,84,88,101-103] and seven reported frequency of errors.[7,12,16,27,35,37,38] Of the research papers, most considered recall and precision. Twelve dealt with recall only,[10,11,62,70,72,76,82,83,90,95,98,100] four considered precision only,[75,78,79,87] and six dealt with time or cost.[20,35,38,84,91,94] All but one of these[20] also considered precision. Other outcomes considered were searches resulting in no hits,[15] novelty,[67] consensus regarding term selection,[30] specificity,[39,88,89] accuracy,[48] relevance,[14] and diagnostic odds ratio.[89]

In analyzing differences in retrieval rates, Dickersin *et al.* classified the causes of retrieval failure as limited use of subject matter MeSH terms; limited use of methodological controlled vocabulary (subject headings), check tags, and publication types]; limited use of free-text subject terms; limited use of free-text methodological terms; and limited use of truncation. Limiting subject matter MeSH terms too severely was thought to be the main cause of retrieval failure.[81]

Several studies explore the role of subject headings in improving recall. Testing four sequential broader search scopes, the only addition that improved recall while preserving precision was the addition of controlled vocabulary.[71] In a search for quality improvement evidence, subject headings showed significantly higher recall than did text words (0.58 versus 0.11, *P* <0.001) although precision was similar (0.26 versus 0.33, *P* = 0.15), and optimal retrieval was seen when subject headings and free-text terms were used together.[74] In a study involving searches for sleep in healthy individuals, parallel subject heading and text word searches were conducted. While the text word only strategy showed higher sensitivity, specificity was higher in the search that used only subject headings.[39]

Jadad and McQuay illustrate how the addition of terms (subject headings and truncated free-text terms) improved MEDLINE recall from 0.67 to 0.90 in a search designed to identify the evidence base for a systematic review.[98] Gotzsche and Lange also reported on the expansion of a search, primarily by including additional methodological terms, to increase MEDLINE recall from 0.93 to 0.98 with a decline in precision from 0.19 to 0.17.[99]

Several studies examine the success factors of searchers with various degrees of training and experience. The use of subject headings contributed to improved recall in most cases. In analyzing reasons for the poorer search performance of experienced end users compared with librarians (end users showed lower recall and lower precision), McKibbon *et al.* noted a greater reliance on MeSH terms and less reliance on text words by the librarians. Librarians also used more advanced features. In a cost comparison, librarians had mean costs that were significantly lower than novice end users, but not lower than experienced end users, and there were no significant differences in search time.[46]

In searches by medical students that failed to retrieve any articles, 28% were judged to have failed because an easily identifiable MeSH term was missed. In a further 10% a less obvious MeSH term was available but not used. These two types were the most frequent errors. Poorly chosen terms (such as using “cardiology” instead of “heart disease”) accounted for another 6% of no-hit searches.[12] Another study of searches conducted by third-year medical students found that more than one-third of missed opportunities involved a missed subject heading.[30]

Markey *et al.* showed that experienced searchers of Education Resources Information Centre (ERIC) tended to miss suitable subject headings, relying instead on free text at the expense of larger retrievals.[85] They found that only 5% of searches in their test set could not be represented through subject headings.

Kirby and Miller analyzed failed searches undertaken by untrained searchers and found that about two-thirds of failed searches were due to missed terms, including subject headings.[10]

There is consistent evidence that subject headings improve recall. There are also several papers that provide insights into the subtleties of effective subject heading use. Sievert and Boyce examined the situation where there were multiple, closely related subject headings and found that recall could be improved by searching single truncated words (rather than the full multiword subject headings) in those subject headings rather than forming a filter by searching all the related subject headings. They describe this reduction of controlled vocabulary as “trimming the hedge.”[84] In another study that explored query expansion from an initial text search based on users’ requests, the addition of thesaurus terms increased recall.[27] A subsequent expansion to related terms further increased recall but with a reduction in precision.

In a study on indexing consistency Funk and Reid showed that indexing consistency was highest in “Anatomy, Organisms and Chemicals and Drugs” branches of MeSH, and lower in the “Analytical, Diagnostic and Therapeutic Techniques” and “Equipment,” “Psychiatry and Psychology,” “Physical Sciences,” and “Health Care” branches. Indexing consistency was higher when the term was central to the topic of the paper and higher for main headings than subheadings.[100] The ability to achieve high recall with good precision may be greatest in the more consistently indexed areas while the less consistent areas may require backup MeSH and free-text terms for high recall, at the expense of precision. This study looks at indexing consistency, not accuracy,[102] and has not been replicated although it is more than 20 years old. A more recent study showed similar levels of consistency in PsycINFO.[104]

a) *Subheadings and floating subheadings*

Subheadings can be used to qualify a specific subject heading, on their own, or as unbound or floating subheadings. Used with a subject heading, they can be expected to increase precision at the expensive of recall. Wright *et al.* limited a central MeSH term in the search to the subheading “methods” (mt). Alone, this limit increased precision from 0.49 to 0.57 and decreased recall from 0.99 to 0.62. When the subheading was used with a search for major MeSH headings only, the precision increased from 0.49 to 0.68 (not a statistically significant improvement) while recall declined from 0.99 to 0.56 (significant at $P < 0.0001$).[92] Dickersin, Scherer, and Lefebvre speculated that the use of “bound” subheadings was one restriction contributing to the poor recall of searches for randomized controlled trials (RCTs) in MEDLINE.[81]

Floating subheadings retrieve all instances of a subheading used with any MeSH term applied to the record. Floating subheadings expand retrieval rather than reducing it, so their use should not be thought of as a limit.[72] Subheading pre-explosion serves to group subheadings that relate to the clinical category being studied, e.g., the pre-explosion subheading “therapeutic use” includes the subheadings “administration and dosage,” “adverse effects,” “contraindication,” and “poisoning” in addition to the subheading “therapeutic use.” In searches for sound clinical studies on diagnosis and treatment pre-exploded subheadings yielded the highest sensitivity, but with decreased precision.[97]

MeSH indexing rules give preference to the application of subheadings instead of equivalent main headings, and searching subheadings as floating subheadings gives larger retrievals.[72] Relying on a subheading instead of including the related subject heading can reduce recall.[67]

Failure to explode subject headings was found to have the greatest impact on the recall of any search errors studied in a group of MEDLINE end users, causing searchers to miss approximately 30% of the references that they wanted.[37]

The hierarchical arrangement of MeSH and other thesauri can be a pitfall. Mismatch on the level of specificity of searching and indexing can lead to retrieval failure.[82]

Even relatively frequent student searchers show confusion over the operation of a thesaurus.[58,78] Exploding may not be understood by some end users who find it logical that they should get everything on specific anesthetics when they use the term "ANESTHETICS," for example.[37]

Allowing automatic term mapping rather than using the MeSH terms with the associated ability to explode broad subject headings had no impact on recall in one effort to develop a search strategy for observational studies. In the same study, the use of major MeSH terms limited as major headings resulted in a decline in recall.[96]

Federiuk found that abbreviations map well to MeSH terms in the OVID search interface but cautions that spacing (e.g., "t-PA" works better than "tPA" to identify "tissue plasminogen activator") and capitalization (e.g., "ACLS" maps to "Advanced Cardiac Life Support" but "acls" maps to "anterior cruciate ligament") matter.[95]

While this review focuses on searchers' errors and database errors, spelling errors do occur and may be a barrier to effective retrieval. Searches of subject headings help guard against misspelled terms in titles or abstracts.[34,70]

b) Identifying subject headings

There were several studies and descriptions of how searchers generate terms.[8,25,26,45,66,78,79,95] The most successful searchers are described as probing the language structure of the database by displaying the indexing of known relevant items or of items highly likely to be relevant and from the displays, picking the terms that will be on target to give good recall and precision simultaneously.[23] Deacon *et al.* examined two approaches — exploration of the MeSH thesaurus with terms suggested by the person requesting the search as a starting point and a strategy where the indexing terms were assigned to known relevant items. The first method produced larger retrievals with higher recall and lower precision than did the second strategy, and required more time and effort to develop.[45] Harden *et al.* also reported on constructing search strategies based on terms used to index known relevant items, but with greater success[20] than Deacon *et al.* Srinivasan used a strategy whereby MeSH terms were selected from documents that ranked high in an initial natural language search, a process that could be conducted automatically.[79] This practice of identifying terms by looking at the indexing of known relevant items has been called "inverse searching."[55]

Fidel described the approach to term selection used by "conceptualist" searchers, who aimed for comprehensive recall. They typically looked at a thesaurus from several angles, and gave thorough treatment to the primary facet in the search question.[8] On the other hand, inexperienced searchers almost never used the index or thesaurus,[16,30] while moderately experienced end users made occasional use of these tools.[16] The most common error was not using MeSH. Less common errors were using MeSH terms that should not have been used or using a term that was too broad or too narrow.[30]

Finding suitable subject headings is a skill that appears to be amenable to training. In one evaluation of an end-user training module, determining specific terms for a search strategy seemed to be the skill that students learned most thoroughly.[17] In another post-training evaluation, residents in a neonatal intensive care setting reported greater confidence in their ability to identify MeSH terms, although their actual ability do so was not reported.[7]

c) Consistency in term selection

McKibbon *et al.* in a study involving three librarians, noted large differences in searching style and use of features but these variations did not result in important differences in precision and recall in a low recall situation (median recall by librarians ranged from 0.37 to 0.47).[46]

Even among experienced searchers there seemed to be little overlap in term selection and in retrievals, but successful searches seem to owe their success to proper selection of terms.[23]

Saracevic and Kantor also reported little overlap in the selection of terms by searchers searching the same question. The more often an article was retrieved by different searchers, however, the greater its odds of being relevant. The recall of these searches was low by the standards of systematic reviewers, in the range of 0.18 to 0.32, with precision in the range of 0.35 to 0.65. They noted similarities between the level of overlap of terms that they found, findings of overlap in searching behaviour in other contexts, and the consistency of terms assigned by indexers. They concluded “the degree of agreement or overlap in human decisions related to representing, searching and retrieving of information is relatively low — the agreement hardly reaches about one-quarter or one-third of the cases involved.”[14]

In a mailed survey, the average agreement among 22 librarians on appropriate MeSH terms for defined concepts was 64% for interventions and 57% for effects. There were terms for which some of those surveyed thought that there were no current acceptable MeSH terms, and if forced to recommend a term, agreement may have declined. Recall using the single most commonly recommended term ranged from 0.27 to 0.86 for interventions and 0.0 to 0.65 for effect variables.[74]

The opportunity to identify missed terms was one of the greatest reported benefits of one peer review effort.[38] Elements covered in peer review included topical vocabulary confirmed through scanning the scope notes of terms and adherence to indexing policies of the selected databases.[24]

Survey respondents saw this element as one of the most important, with the potential for a large impact on precision and recall (Table 4). Errors here were seen as a mark of inexperience. The importance of missed subject headings was rated as having strong research support as well as being self-evident.

6.2 Second Tier Elements

6.2.1 Natural language terms missing

Seventy-one papers included in the review addressed the effect of natural language search terms on retrieval. Fifty-one were research-based, and 10 dealt with the frequency of errors. Nineteen provided discussion but no empirical evidence, and so are not addressed here.

Of the research papers, 46 dealt with recall,[7,8,10-12,14,16,20,23,24,27,34,36-39,41,45,48,59,62,64,66,67,69,73,74,76,77,80,81,84-86,88-91,93-95,97-100,105] 34 addressed precision,[7,8,12,14,16,23,27,34,36-38,41,45,48,59,66,67,69,73,74,77,78,80,81,84-86,88,89,91,93,94,97,99] seven reported specificity,[39,48,66,88,89,97,105] six mentioned time and cost,[20,36,38,84,91,94] and six reported other parameters.

Some studies looking at search effectiveness do not distinguish between the topics of natural language terms and subject headings in their results.[10,17,98] Because many recommendations are shared between these topics, we do not restate the finding unless they have additional implications for the peer review of natural language terms.

Savoy found that searching the abstract, in addition to the title and subject headings, increased search performance over searching only the title and subject headings.[86] Similarly, Adams *et al.* attribute the improvement in sensitivity over time of a search for RCTs in mental health to the introduction of abstracts to MEDLINE in 1976.[91]

Watson and Richardson evaluated relevant but unretrieved records and found that missed natural language variants describing group therapy (for example, group focused, group sessions, and group interventions) accounted for many misses. Further misses occurred in PsycINFO when RCTs were not indexed with terms related to study design as they had been in MEDLINE. The searchers, however, had not put in any free-text statements to capture the methodological aspect, although the methodology was stated in the abstract.[59] McKinin *et al.* identified five classes of problems, "(1) those occurring because the author used a variant form of a term; (2) those occurring because the author used a more general term than the searcher; (3) those occurring because the author used an abbreviation, acronym or chemical formula; (4) those occurring because of the author's use of a synonym not included in the searcher's hedge; and (5) those occurring because the author referred to a member of the class not included in the searcher's hedge." Natural language problems accounted for 33% of retrieval failures in full-text searching.[11]

Sewell and Teitelbaum found that 22% of problems were with natural language, usually missed synonyms, although these had a low impact on search performance relative to that of errors in the use of MeSH terms.[37] An analysis by Wildemuth and Moore of end-user searching errors made by third-year medical students found that although a student's initial selection of terms was adequate, missed opportunities included missed synonyms.[30] In a review of searches for RCTs, Dickersin, Scherer, and Lefebvre found that the limited use of text word searching was the most consistent defect.[81]

Markey *et al.* while studying searches of the Education Resources Information Centre (ERIC) database, found that for 5% of searches, the concepts could not be adequately translated into subject headings, because suitable subject headings did not exist. They suggested six categories of searches that might be better served by natural language queries: geographical areas, recent topics, specific named objects, value judgments (adjectives), actions statements (verbs), and individual or psychological characteristics (adjectives).[85]

Survey opinion placed “missing natural language terms” as a second tier element because of its importance in peer review with the potential for a large impact on search performance, but ranked this element behind subject headings in importance (Table 4).

6.2.2 Subject headings and natural language terms are needed

Rowley stated in 1994: “There is general recognition that controlled language and natural language should be used in conjunction with one another…This is based, however, on practice and experience rather than proved and tested research.”[106] We have identified a body of evidence that shows that optimal retrieval is achieved through the use of a combination of subject headings and natural language terms.

In a systematic review of studies comparing ≥2 search strategies for randomized trials in MEDLINE, Crumley found that overall, searches using keywords and MeSH had better recall and precision.[94]

Watson and Richardson expanded searches for three databases that were initially made up of subject headings to include free-text terms. Recall improved from 0.84 to 0.97 in MEDLINE, 0.68 to 0.76 in EMBASE, and 0.38 to 0.65 in PsycINFO. Precision declined from 0.57 to 0.35 in MEDLINE, 0.48 to 0.37 in EMBASE, and 0.47 to 0.39 in PsycINFO.[69]

van der Weijen *et al.* examined MeSH terms and natural language terms for two diagnostic tests and found that the addition of the natural language search increased recall with a slight loss of precision.[90]

Validated search filters, mostly methodological filters, almost always achieve optimal performance (usually high recall with some preservation of precision) through a combination of subject headings and natural language search terms.[48,73,74,81,87,89] The retrieval of clinically important studies in MEDLINE can be enhanced by combinations of indexing terms and text words.[48] The authors systematically tested subject headings and text words to construct a search filter for diagnostic test evaluations. The most accurate strategy (recall 0.80, precision 0.48) contained natural language terms and subject headings.[89]

Although the main comparison in McKinin *et al.*’s study was between MEDLINE and full-text searching, a subset of topics was searched twice, once with subject headings only and once with natural language with or without subject headings (at the searcher’s discretion). The natural language only and subject heading only searches showed no difference in recall or precision.[11]

Muddamalle undertook 81 searches on topics related to soil mechanics, testing natural language and thesaurus-based searches for each. While both types of searches yielded similar results for recall and precision, the combination of both exceeded the performance of either, increasing recall by 5%.[77]

Across 11 topics and five databases, the recall percentage for each data base was improved by the addition of relevant documents found only by a natural language search strategy.[67] Increases ranged from 0.05 added (0.13 increasing to 0.18) to 0.17 added (0.12 to 0.29). The highest recall achieved was with MEDLINE, with the addition of natural language terms increasing retrieval 0.10 from 0.27 to 0.37.

In a study examining indexing and retrieval approaches, Srinivasan found that the most effective indexing strategy is one where independent MeSH and free-text indexes are maintained and are used in combination during retrieval. For optimal retrieval, the free-text terms are selected from the original question, and the MeSH terms are selected from the top-ranked documents retrieved by the initial text word search.[79]

6.2.3 Missed spelling variants and truncation

Spelling variants are an issue in free-text search statements, where suffixes and spelling variants must be addressed in a character-string matching process.[25]

We found 30 eligible reports. Sixteen were research reports of the impact on search performance,[7,10,11,16,23,30,36,38,47,62,66,69,81,88,95,107] six dealt with the frequency of the error,[7,16,30,36,38,108] and 13 were discussions without research evidence.[9,18,19,21,25,26,33,51-55] Of the 16 studies of impact on search performance, 14 identified an impact on recall, and 9 also considered precision. One also considered time and cost and discussed peer review as a means of addressing errors in this area.[38] Two studies seemed to refute the importance of spelling variants and truncation in successful retrieval.[10,23]

Watson and Richardson, in a study involving MEDLINE, EMBASE, and PsycINFO, found that free-text terms and truncated vocabulary improved sensitivity in all three database studies. In all cases the gain was at the expected cost to precision.[69]

In a study of success factors in untrained end users, truncation was found more often in failed searches than successful searches. Failed searches did not retrieve relevant references. These results run counter to conventional wisdom but are internally consistent – search manoeuvres usually associated with precision rather than recall were seen more often in searches that successfully retrieved relevant material. Failed searches used the Boolean operator "OR," a manoeuvre that usually increases recall more often than successful searches. The authors interpret this result to mean that the searches failed because although the searchers seemed to recognize a need for different terms and approaches, they were unable to find effective ones.[10]

In a study of experienced searchers, Bellardo and Saracevic found little or no difference in the general use of truncation and search devices generally recommended to improve recall or precision. Rather, the presence of fatal errors distinguished poor searches, and a better selection of search terms established through a refinement process seemed to distinguish good searches.[23]

Of "missed opportunities" documented by Wildemuth and Moore, truncation was the eighth most common omission (of 14 types), accounting for about 10% of missed opportunities.[30] McKinin *et al.* examined reasons for retrieval failure in full-text articles and found problems of word variants in 5% of cases.[11] Thirty-one per cent of search strategies by novice searchers studied by Bysouth used truncation. The proper use of truncation could have improved 39% of the novices' flawed strategies, and inadvisable or incomplete truncation was seen in 19% of basically flawed searches.[36]

Several search assessment tools examine whether the searcher has addressed spelling variants and truncation,[7,16,18,21,30,47,52] and these elements are often discussed in search technique instruction.[26,54] Rigorously developed search strategies such as methodological filters generally include truncation.[66,81,88]

Several authors provide examples of missed variants.[38,62] In one example involving a search for material related to fluorodeoxyglucose (FDG), 56 spelling variants for FDG were identified, although complete retrieval was accomplished with 11 well-chosen variants and truncated terms.[107] As well as spelling variants, variants of punctuation occurring in users' queries include possessive forms with an apostrophe, acronyms with periods between letters, and hyphenated words and phrases.[29] Systems differ in their handling of internal punctuation, providing an additional pitfall for those without extensive experience in a particular search interface. Another complexity is variation in internal hyphenation in an abbreviation — for instance, t-PA and tPA.[95] These examples cannot be addressed through truncation.

Some search strategies incorporate common misspelling to capture incorrect spelling variants that may be present in the records. An example is the search strategy to identify systematic reviews in CINAHL, which includes the correct and incorrect spellings for the PsycINFO database in line 12 (available at http://www.york.ac.uk/inst/crd/search.htm#CINAHL, visited September 6, 2007).

This area was addressed through two survey questions, "Are any spelling variants missing?" and "Are there any errors in truncation?" While most survey respondents supported their inclusion in peer review, these elements were not seen as having the consistently high impact of first tier elements (Table 4).

6.2.4 Irrelevant subject headings

Three papers discussed irrelevant subject headings. Two research papers addressed the impact on precision.[78,81] One additional study provided indirect evidence on precision.[7] One paper[78] looked at the frequency of error.

Survey respondents ranked irrelevant subject headings as potentially having a large negative impact on precision (Table 4). This element, with failure to adapt the search for different databases, errors in line numbers, and errors in Boolean operators, was rated as having the greatest impact on precision. The element "Irrelevant subject headings" was in the second tier of importance for peer review.

6.2.5 Irrelevant natural language terms

Four papers examined irrelevant natural language terms. One paper examined the frequency of errors and their impact on precision.[78] One considered this as an aspect of peer review,[38] and two provided a theoretical rationale or discussion of the issue.[7,25]

Spink and Saracevic found that more than one-third of all terms used in mediated online searching produced nothing but irrelevant answers or had no retrievals at all.[78] One published example of a peer review of a literature search identified two terms included in a search on hemophilia that were not relevant to the question.[38]

In a review of online search development techniques, Hawking and Wagers recommend formulating alternative strategies, reviewing results, and assessing them against previous attempts. "One then continues to alter the search, adding synonyms and excluding less valuable terms, until the final result is reached."(p. 13). They caution against accepting a list of keywords from the requestor rather than a narrative statement of the search topic.[25]

"Including irrelevant natural language terms" fell in the middle tier of importance. Survey respondents assessed the inclusion of irrelevant natural language terms as predominantly having a negative impact on precision and of being moderately important in peer review (Table 4).

6.2.6 Limits

Limits, by narrowing the search, are used to increase precision. Given the emphasis on high recall, these must be used carefully in systematic review searches. The main limits considered here are subheadings, subject headings restricted to major focus only,[92,96] methodological filters or hedges, and tags such as age, human, language, and study type.

Thirty-one articles in this review addressed limits. Thirteen articles provided research evidence.[13,23,39,45,48,49,66,67,72,81,92,97,100] Of these, 11 considered recall,[13,19,23,39,48,66,67,92,97,100,109] five considered precision,[23,48,66,92,97] and three considered specificity.[48,66,97] Two papers reported on the frequency of errors,[30,36] and 15 provided a theoretical rationale for the impact, or provided search instruction or advice.[7,16,19,22,24,28,33,47,54,55,58,101,103,110,111]

Using Major MeSH or restrict to focus (central concept) main headings, searched through "starring" a MeSH term (with equivalents in other databases) may increase precision at the expense of recall as not all records that are relevant to the review may have the term as a major subject heading. Wright *et al.* limited 2 searches this way and saw precision change from 0.49 to 0.50 in one search and 0.28 to 0.40 in another while recall changed from 0.99 to 0.84 in one search and from 1.00 to 0.42 in the other. Declines in recall were statistically significant.[92]

Subheadings that are used to qualify subject headings could be considered as limits but are discussed under the "subject headings" element. In most instances, indexers typically use no more than three subject headings per index term.

MEDLINE and many other databases provide additional indexing that enables searchers to limit the search by age, human, year or language of publication, or to certain publication forms or types. These additional indexing features vary by database. Two potential problems with tags lie on the indexing side. The first issue is that tags, when assigned, must be accurate. Funk and Reid's work on indexing consistency found that check tags were applied more consistently than MeSH or subheadings, and concluded that they could be used reliably to limit retrieval. Yet consistency was only 74.7%, indicating some inaccuracy or incomplete tagging. The tags may not be assigned in all cases where they would be relevant.[100] Deacon, Smith, and Tow documented that almost all indexers missed age group terms at various times.[45] Recall was lowered by child tags when a simple "NOT" was used.[39] Thus, such limits need to be applied through the careful use of "NOT," so that unwanted tags are excluded but records with no tags on the dimension of interest remain. There must also be provision for multiple tagging, where one of the tags applied is the desired category. This results in the construction "not animal (not animal and human)", which is used to exclude animal studies that do not involve humans and animals, rather than limiting to the desired "humans" — a simpler construction but one that would exclude studies not tagged as either but that included relevant human evidence.

In Medline, McKinin *et al.* found that limits by language of publication, date, and Core Clinical Journals (formerly Abridged Index Medicus or AIM) were effective at improving precision in searches addressing standard use queries (but not systematic reviews).[13]

Tenopir and Shu found that, in full-text searching, precision could be improved by excluding certain publication types defined by the user as irrelevant.[49] The Hedges team applied "Publication Type" limits in a preliminary screening step to increase precision.[48] Editorials, comments, letters, and news were excluded from the search using the Boolean "AND NOT" operator.

Methodological filters or hedges (including PubMed and MEDLINE clinical queries) have been reviewed by Jenkins,[66] who concluded that "relying on the performance measures alone to make adequate judgments about the applicability and validity of search filters is not sufficient. In order to make informed decisions on the use of search filters, an awareness of the limitations of the methodology used in the development of such search aids is required."[66]

One examination of novice search errors identified the non-use of subheadings and the failure to limit a term to major subject headings as frequent missed opportunities.[30] In a study of information retrieval by research scientists, the limits considered are review, year, book, human, and corporate source — and errors were infrequent.[36] One study comparing successful moderate recall searches and less successful searches found little or no difference in the use of limits such as language, date, or publication types.[23]

Apart from their effect on recall and precision, limits not relevant to the question should be flagged for evaluation of the potential for bias. When they are used to limit the search based on factors not relevant to the question (for instance, to exclude non-English language publications or certain publication types such as dissertations or conference abstracts that could be considered to be grey literature), they can introduce bias into the review. A review of the literature on epidemiological bias is beyond this study's scope but Felson[112] examined the bias that can be introduced into a systematic review at various stages, and Egger and Smith review bias in the location and selection of studies for meta-analysis.[113]

Survey respondents rated unwarranted limits as a greater problem than missing limits, because the excessive limits were seen as having a large negative impact on recall while the impact of missing limits was on precision (Table 4). Survey respondents considered limits to be an important element in peer reviewing.

6.3 Third-Tier Elements

6.3.1 Additional fields

Most searching is based on subject headings and natural language in the title and abstract of the bibliographic record.[114] Databases, however, often contain other fields that may be useful to the search, e.g., drug codes, publication types, authors, journal titles, or addresses.

Thirteen reports dealt with special fields, but usually only in passing. Six research papers dealt with recall,[23,67,71,73,91,115] and most of these also considered the effect on precision. Six additional papers discussed special fields.[16,22,25,52,53] One reported on the use of fields by research scientists after training.[36]

Additional fields that can be searched include author, publication type, CAS registry numbers (unique numerical identifiers for chemical compounds, polymers, biological sequences, mixtures and alloys), or other drug indexes. EMBASE allows extending searches to retrieve from section headings.

BIOSIS has Concept Codes, and INSPEC has section codes. Enhanced recall is the main search parameter influenced by searching additional fields.

"Additional fields" is a third tier search element. Evidence is sporadic, but there are several special fields that can be tapped into for enhanced retrieval. Survey respondents did not see this as a critical aspect of the search to focus on in peer review (Table 4).

6.3.2 Organization of search and search logic

We hypothesized that long or complex search strategies, such as those used in systematic reviews, would be more difficult to peer review and might be more prone to error if they were poorly organized.

Six papers discussed search logic and organization.[16,22,24,25,28,54] One of these presented research findings.[16] One noted a potential time and cost saving by searching the most specific aspect.[25] One described how this aspect was dealt with in peer review: "The goal of the reviewer is to check whether the searcher has identified the components of the request and separated them in logical groups."[24] Although the evidence provided by any one article is sparse, the papers in this set are influential. They include the Cochrane Handbook, the CASPfew site, and the work of Reva Basch, the paper that was the most complete search checklist we found and the description of the most complete peer review effort we found.

The only research report identified did not find an association between search clarity and retrieval effectiveness.[16] No data on prevalence of problems in search logic or organization were found.

An authoritative discussion of the logically organized and presented search is presented by Hawkins and Wagers,[25] who rely on Buntrock's work.[116] One approach to searching is the building block method, where concepts are built and combined, as is often done when searching questions formulated according to the PICO model. Hawkins and Wagers discuss the advantages of such an approach, chiefly that the logical organization makes the search easy to follow and to review and understand, even after the passage of time. These are advantages for peer review and for updating, which is an important aspect of systematic reviews.[25] The disadvantage of such a rigid structure is that it is seen as impairing creativity and the ability to pursue unforeseen opportunities. The authors describe more flexible and creative approaches that could be useful as exploratory techniques during the development of systematic review searches.

The Cochrane Handbook recommends a similar approach: "An electronic search strategy should generally have three sets of terms: 1) terms to search for the health condition of interest; 2) terms to search for the intervention(s) evaluated; and 3) terms to search for the types of study design to be included (typically randomized trials)." And "a good approach to developing an electronic search strategy is to begin with multiple terms that describe the health condition of interest and join these together with the Boolean 'OR' operator. This means you will retrieve articles containing at least one of these search terms. You can do likewise for a second set of terms related to the intervention(s) and for a third set of terms related to the appropriate study design. These three sets of terms can then be joined together with the 'AND' operator. This final step of joining the three sets with the 'AND' operator limits the retrieved set to articles of the appropriate study design that address both the health condition of interest and the intervention(s) to be evaluated."[28]

One searching guide recommended organizing the search so that limits were placed at the end.[54] Interviews with “super searchers” regarding how they plan their search strategies indicated at least some planning, but with variation in how formally this was done, varying from quick and dirty to a more formal building block approach. These searchers worked predominately in the business environment rather than health research.[22]

Based on the review of the evidence, this element was divided into two survey questions, the first addressing translation of the question into search concepts and the second addressing search organization (Appendix C questions 1 and 2). Conceptualization is reported under a separate heading.

Survey respondents placed little importance on the organization of the search, which many saw as having moderate impact on recall and little impact on precision (Table 4). A poorly organized search was not seen as indicative of unfamiliarity with aspects of searching.

6.3.3 Redundancies

We specified redundancy “without rationale” because some searchers may, for example, report a narrower term and an exploded broader term to clarify to readers who are unfamiliar with subject heading hierarchy that both terms factor into the search.

Two papers addressed this element. Jokic identified redundancy as having no direct impact on the search result but as having an influence on search efficiency and grouped this error with others such as browsing already seen references. This type of error was taken to be a mark of inexperience, specifically a lack of knowledge of search systems and syntax. The error was seen as having no direct influence on search results but as influencing search efficiency.[16] Goodman cautions systematic reviewers that redundancy may be useful to protect against indexing inconsistencies and errors and provides the example of searching MEDLINE for reports of RCTs where “it may be useful to use the search command: RANDOMIZED CONTROLLED TRIAL (PT) OR RANDOMIZED CONTROLLED TRIALS because an indexer may have used the latter term in place of the first.”[55]

Survey respondents did not expect this element to affect recall or precision. They did not see it as a notable indicator of unfamiliarity with searching and did not expect much research evidence on this. There was, however, some support for peer reviewing for redundancies, although it fell in the third tier of importance (Table 4).

6.3.4 Subject headings and natural language terms combined

We expected that an approach to searching that involved free text and subject headings would provide the strongest performance, but we also anticipated that best practice would be to separate them on different lines of the search (i.e., as different search statements, that would later be joined by the Boolean “OR”).

No evidence on the impact or prevalence of combining subject headings and natural language terms in the same line was found.

Survey respondents assessed this as having a potential impact on recall, but none on precision. They did not believe that there was a strong basis of evidence for this element nor was it given importance as an aspect of the search for peer review, although there was agreement that it should be considered.

6.3.5 Subject headings exploded with no narrower terms

We hypothesized that exploding subject headings with no narrower terms might be a sign of poor understanding of the workings of a thesaurus and might be an indication of carelessness in exploring the thesaurus, thus being a marker for other potential problems in the search strategy. One case study of peer reviewing search strategies addressed this element and marked it as an error in the search.[38]

While survey respondents agreed that exploding a subject heading that had no narrower terms may be a sign of unfamiliarity with the workings of thesauri, it was not ranked high in importance, relative to other elements (Table 4).

7 PEER REVIEW FORUM

7.1 Peer Review Forum Pilot

Before programming the interface for a fully operational version of the peer review forum, 10 survey respondents participated in a pilot of a web-based peer review forum in which 10 sample searches selected from The Cochrane Library, published journal articles, and HTA searches (Appendix F) were assessed using the 7 point PRESS checklist (Appendix G).

7.1.1 Methods

An SRS page (with TrialStat!) was arranged with the sample searches. We used 10 searches of different lengths and complexities and with a range of errors. The clinical question was posted with the search so that the participants could see the actual question and rate how well it was translated. The search strategies were presented in their original format. At least one search had no discernable errors. We introduced additional errors. We then determined and agreed on the errors in each search that were to be used for comparison. Each item was scored as “satisfactory” or “needs revision,” and the overall search was classified as “satisfactory” or “needs revision.”

a) Recruitment

Librarians (participants) who responded to the survey were then invited to participate as junior or senior according to their survey responses. The junior participant group included five librarians with ≤5 years of database searching experience and who had undertaken at least one but not more than 10 systematic review searches. The senior participant group included five librarians with ≥5 years of searching experience and experience with ≥20 systematic review searches.

b) Training

Participants made a practice rating of one search and were given the opportunity to ask questions and revise their ratings. Questions and answers were shared with all participants by email, but the identity of participants was concealed.

c) Validation ratings

Participants were asked to complete search ratings independently, without discussion. Participants were asked to complete the 10 assessments within two weeks, spending no more than 30 minutes assessing each search. The rating scale was presented in HTML format with hyperlinks to supplemental instructions and examples to keep the form of presentation similar to what it would be in the forum. Cochrane's Q was used to examine the homogeneity or equality of rate of "fault finding" by forum participants for each element of the PRESS forum. The significance level was set at 0.01 to correct for multiple testing.

7.1.2 Results

For all elements except limits, forum participants rated fewer elements as "satisfactory" than did the project investigators. Across the 10 searches, on average, forum participants rated 3.3 of seven elements adequate whereas investigators rated 5.3 elements as adequate. While there was variability in the responses on all elements, there was significant between-rater heterogeneity for three elements. These elements were the choice of subject headings, natural language terms, and limits. For example, of the 10 searches rated, two participants assessed all 10 as inadequate on subject heading selection, and three rated 9/10 as inadequate on subject headings. One participant rated 3/10 searches as inadequate, and another rated 4/10 as inadequate. For the three elements with significantly heterogeneous rates, there was no discernable difference between the number of searches assessed as inadequate by junior and senior participants.

7.1.3 Discussion

While peer review forum participants judged more elements as inadequate than did the investigators, they rated four of seven elements consistently. Two of the elements showing inconsistency involved choice of terms: subject headings and natural language terms. These were also the two elements least often assessed as adequate by the peer review forum participants. The choice of terms is known to vary between searchers,[23,46,74] but core documents seem to be identified in most searches, regardless of the exact terms selected.[23,46] All the research that has highlighted variability in term selection has been done in contexts other than systematic review searches. Systematic review searches are distinct in that they require high recall, and the effect of variability in term selection has not been studied in this context, although it seems to be an important area for investigation. Thus, although the potential lack of agreement on term selection could become a problem in peer review, the ability to have a second information specialist review terms and propose alternatives may also be of benefit. Participants did not have the opportunity to read the systematic review of evidence regarding search error before participating in the exercise, although guidance derived from the review was presented with the elements on the forum page. The participants' assessments may have more closely matched those of the project investigators had they been able to access the systematic review before participation in the forum.

Procedurally, the application of the PRESS checklist (Appendix G) to systematic review searches is viable for junior and senior librarians. An educational intervention is probably required to train librarians in using the PRESS checklist if a high level of consistency in the more subjective aspects is sought.

7.2 Development of Forum Interface

A web-based forum was developed as an environment for posting and reviewing search strategies. The Peer Review Forum is a network of search experts with a common interest in the peer review of electronic search strategies using the PRESS checklist during the initial stage of a systematic review or HTA.

The purpose of the peer review forum is to evaluate the search strategies of peers, as opposed to search strategy training or the writing of search strategies. This may result in improved search quality and will provide a method for the searching aspect of a systematic review or HTA that will be on par with the rest of the review's methods. This will enhance the review's scientific rigour. Arrangements for peer reviewers to receive continuing education credits for their reviewing efforts will be discussed with health library associations and the Information Retrieval Methods Group of The Cochrane Collaboration.

The PRESS web site was developed by ASP.NET using Microsoft Visual Web Developer and Microsoft SQL Server Express (Appendix H). The navigation page includes left menu and right menu and shows the current selected page and login status.

The PRESS web site provides the following functionalities:

Administration and security
- login
- new account
- new profile
- forgot your password
- change password

Main activities
- put answer
- put question
- create and edit profiles
- view question statement with answers
- send email
- uploading and downloading files.

8 DISCUSSION

8.1 Results

8.1.1 Existing checklists

Twenty-six reports containing aspects that could be used as search assessment checklists were identified. A few of these addressed the conduct or reporting of the larger search used to form the evidence base for systematic reviews, HTA reports, or clinical practice guidelines, but did not address the quality of the electronic search strategy. These have been reported elsewhere.[4] Additional checklists were developed to assess student learning (Appendix D). Few of these are validated. No examples of validated checklists for assessing search strategies were found. Many of these checklists did inform the construction of the PRESS checklist as they also reported research data on the impact or prevalence of search errors.

8.1.2 Evidence regarding search elements

Elements with enough evidence to support their use in the peer review of electronic search strategies are:

- conceptualization of research question
- spelling errors and wrong line numbers
- translation of search strategy to different databases
- missed subject headings
- missed natural language search terms
- spelling variants and truncation
- irrelevant subject headings
- irrelevant natural language terms
- search limits.

There is research evidence to support the view that problems in these elements will have a negative impact on search performance. Furthermore, there is support from the community of librarians who do information retrieval for systematic reviews or HTAs on the importance of peer reviewing these elements.

The elements vary in the amount of specialized knowledge needed to peer review them. Some elements, such as spelling errors or incorrect line numbers, are largely mechanical and require no special knowledge to assess for accuracy. Others, such as translation of the research question into a series of connected search concepts, require expertise to implement and to assess.

The selection of terms requires specialized knowledge, such as an understanding of the operation of a thesaurus, but also appears, based on the literature, to involve an element of judgement. There is no readily identifiable single "correct" set of terms for implementing the search for a particular question or "one best" method or strategy for searching — sometimes several approaches work equally well.[117] This is made apparent if one considers the rigorous process needed to establish a validated filter.[118] A validated filter is developed in one set of data and then the findings are checked out against another similar set of data to validate the first findings. Even at this extreme, two iterations of a search designed to identify reports of controlled clinical trials, both created using rigorous and reproducible methods, but done 10 years apart, look different.[119-121]

Thus, the peer review of a search will involve an ascertainment that no technical errors have been made and a more subjective assessment of the adequacy of term selection.

It seems unlikely, even undesirable, that the peer review of search strategies can be done as part of the general peer review of a finished research report. First, those reviewers, if they are selected for their knowledge of the topic being studied, may not have the specialized knowledge needed to assess the adequacy of the search. Second, if errors are discovered at that point, after the systematic review is completed, it is more problematic than, for example, it would be if the reviewers requested a change to the statistical tests used. Thus timing and knowledge requirements make peer review more practical at the start of the project.

The early peer review of search strategies can be accomplished in several ways. First, it can be part of the protocol approval process. Few published systematic reviews, other than those produced by The Cochrane Collaboration, state that the systematic review was based on a protocol.[122] The consistent use of approved protocols by The Cochrane Collaboration demonstrates the feasibility of the approach. Larger work groups, like national HTA organizations, may have enough expertise to provide in-house peer review. A third approach is to develop a peer review forum, a community of interested librarians who have the expertise to create and review the type of searches needed in systematic reviews and HTAs. We have undertaken a pilot of such a forum,[123] with participation from librarians associated with Cochrane and with national HTA agencies, and are continuing the process of refining and evaluating this approach.

Early and evidence-based peer review of search strategies may result in improved quality and will provide a method for the searching aspect of a systematic review or HTA report that will be on par with the rest of the review's methods. This will enhance the review's scientific rigour.

8.2 Strengths and Weaknesses of this Assessment

The strengths of this assessment are that research evidence, theory, and experts' opinions are integrated into a comprehensive appraisal of the elements of successful electronic search strategy for systematic review searching.

Limitations are that this assessment addresses only optimal electronic search strategies. Additional decisions will be made and implemented in any systematic review, which must be informed by research that is outside the scope of this assessment. Those decisions include the choice of databases to be searched[124] and additional sources to be consulted.[125] Several survey respondents added comments on the importance of those decisions, although they are beyond the scope of evidence-based peer review of the electronic search strategy and are influenced by the resources, time, and money available to the review team.

This assessment addresses limits, including methodological filters such as sensitive search strategies,[126] or limits by language[127,128] or by publication type.[129] These are rich areas of research, and those responsible for searching the literature for systematic reviews need to be aware of and keep current with those developments.[129]

We did not perform a quality assessment of the research evidence included in this review. Many research methods are used, and we do not have the expertise or appraisal instruments to assess the quality of all types. Thus, some of the evidence may come from studies that are susceptible to bias.

8.3 Generalizability of Findings

These findings and the peer review elements derived from them are largely consistent with current best practice as described in authoritative sources such as The Cochrane Handbook for Systematic Reviews of Interventions.[130] Input was sought from the health science library community, from the HTA information retrieval community, and from The Cochrane Collaboration. The preliminary results of this work have been presented to these communities for feedback.

Research evidence was drawn from all aspects of library science, not just health sciences librarianship. Thus, the peer review elements identified could be applied to any type of literature search, although maximizing recall was given the greatest weight in the formulation of these peer review recommendations, because recall is the most important parameter of search performance for systematic reviews and HTAs. In other situations, other parameters such as precision or cost might be of greater importance, so the focus of the peer review would need to be adjusted accordingly.

Some technical aspects of the searches discussed in this report are database- or interface-specific; for example, floating subheadings are a feature available in MEDLINE, at least through the OVID interface. MEDLINE is the most commonly used general health care database in the systematic reviews of health care interventions, at least those done by The Cochrane Collaboration, and OVID is the interface most commonly reported for searching MEDLINE.[131] Hence, some of these features will be absent in other databases. Conversely, databases that are not commonly used in health science systematic reviews and HTAs may have other features that could be considered in peer review yet are not addressed in this report.

8.4 Knowledge Gaps

Several of the elements examined lacked research evidence of their impact on search performance. In most cases, the survey respondents assessed the impact as self-evident, but several gaps remain. Pertinent to systematic review and HTA, the importance of tailoring searches to individual databases has not been explored. Although there is a large body of evidence suggesting that more than one database needs to be searched for optimum retrieval of relevant evidence, the factors leading to success remain unclear. It may be that the unique coverage of additional databases helps to complete the evidence base, or it may be overlapping coverage but with different indexing that gives the searcher an additional, independent chance to identify a relevant study. Other areas with little or no research evidence that may warrant more research include the impact of truncation errors and of including irrelevant subject headings.

The pilot of the PRESS checklist (Appendix G) showed that some inconsistency in the application of the checklist results when raters assess the searches without study or training with regard to the relevant evidence. An educational intervention is likely needed to train librarians in using the PRESS checklist if a high level of consistency in the more subjective aspects is sought.

9 CONCLUSIONS

This work fills a gap in the quality assurance of search methods in systematic reviews and HTA reports. Errors in the electronic search strategy have been shown to reduce the effectiveness of electronic search strategies used in systematic reviews and HTA reports. Without quality assurance of a bias-free and complete evidence base, the true outcomes of a systematic review cannot be tested. The deliverables from this project include an evidence-based process, developed for the peer review of the electronic search strategy, and consisting of a validated checklist and an electronic peer review forum.

The methods for this evidence-based peer review process were developed after performing a systematic review to identify appropriate literature, a web-based survey, and an interactive peer review forum of search experts. These methods supported the importance of specific elements that were used in developing a checklist and peer review process for evaluating the success of the electronic search strategy.

Those involved in systematic reviews and HTA reports should have an evidence base for their work that is based on a peer reviewed electronic search strategy, because the greatest portion of their evidence base will be created during this search.

The peer review process may delay the development of the electronic search strategy but considerations for this have been built into its development. The trade-off for taking more time will ensure that quality assurance has occurred. The evidence-based peer review of electronic search strategies requires the same body of expert knowledge needed to create search strategies. The peer review should be undertaken by librarians or other suitably qualified and experienced information scientists. The only costs associated with the proposed peer review process pertain to the ongoing hosting and management of a web site. It is proposed that peer reviewers volunteer for this task. Participation is based on the principle of reciprocity. Participants will be able to have their searches peer reviewed in exchange for peer reviewing the searches of others.

10 REFERENCES

1. Sampson M, Tetzlaff J, McGowan J, Cogo E, Moher D. No consensus exists on search reporting methods for systemic reviews [In Press]. J Clin Epidemiol 2008.
2. Sampson M, McGowan J. Errors in electronic search strategies of Cochrane reviews were identified by type and frequency. J Clin Epidemiol 2006; 59:1057-1064.
3. Koufogiannakis D, Slater L, Crumley E. A content analysis of librarianship research. J Inf Sci 2004; 30(3):227-239.
4. Landis JR, Koch GG, Freeman JL, Freeman DHJ, Lehnen RC. A general methodology for the analysis of experiments with repeated measurement of categorical data. Biometrics 1977; 33(1):138-158.
5. Kagolovsky Y, Moehr JR. Current status of the evaluation of information retrieval. J Med Syst 2003; 27(5):409-424.
6. Booth A, O'Rourke AJ, Ford NJ. Structuring the pre-search reference interview: a useful technique for handling clinical questions. Bull Med Libr Assoc 2000; 88(3):239-246.
7. Bradley DR, Rana GK, Martin PW, Schumacher RE. Real-time, evidence-based medicine instruction: a randomized controlled trial in a neonatal intensive care unit. J Med Libr Assoc 2002; 90(2):194-201.
8. Fidel R. Online searching styles: a case-study - based model of searching behaviour. J Am Soc Inf Sci 1984; 35(4):211-221.
9. Fidel R. Searchers' selection of search keys: 3. Searching styles. J Am Soc Inf Sci 1991; 42(7):515-527.
10. Kirby M, Miller N. MEDLINE searching on colleague: reasons for failure or success of untrained end users. Med Ref Serv Q 1986; 5(3):17-34.
11. McKinin EJ, Sievert M, Johnson ED, Mitchell JA. The Medline/full-text research project. J Am Soc Inf Sci 1991; 42(4):297-307.
12. Mitchell JA, Johnson ED, Hewett JE, Proud VK. Medical students using Grateful Med: analysis of failed searches and a six-month follow-up study. Comput Biomed Res 1992; 25:43-55.
13. Saracevic T, Kantor P. A study of information seeking and retrieving. 2. Users, questions, and effectiveness. J Am Soc Inf Sci 1988; 39(3):177-196.
14. Saracevic T, Kantor P. A study of information seeking and retrieving. 3. Searchers, searches, and overlap. J Am Soc Inf Sci 1988; 39(3):197-216.
15. Walker CJ, McKibbon AK, Hayes BR, Ramsden MF. Problems encountered by clinical end users of MEDLINE and GRATEFUL MED. Bull Med Libr Assoc 1991; 79(1):67-69.
16. Jokic M. Analysis of users' searches of CD-ROM databases in the national and university library in Zagreb. Inf Process Manag 1997; 33(6):785-802.

17. Mularski CA, Bradigan PS. End-user searching: review of a modular program. Bull Med Libr Assoc 1993; 81(1):61-63.

18. Nahl D. Affective elaborations in Boolean search instructions for novices: effects on comprehension, self-confidence, and error type. In: *Proceedings of the 58th Annual Meeting of the American Society for Information Science; 1995*.

19. Allison JJ, Kiefe CI, Weissman NW, Carter J, Centor RM. The art and science of searching MEDLINE to answer clinical questions: finding the right number of articles. Int J Technol Assess Health Care 1999; 15(2):281-296.

20. Harden A, Peersman G, Oliver S, Oakley A. Identifying primary research on electronic databases to inform decision-making in health promotion: the case of sexual health promotion. Health Educ J 1999; 58(3):290-301.

21. Association of College and Research Libraries. *Information literacy competency standards for higher education*. Chicago: American Library Association, 2000.

22. Basch R. Secrets of the super searchers: planning search strategies. Online 1993; 17(5):52-58.

23. Bellardo T, Saracevic T. Online searching and search output: relationships between overlap, relevance, recall and precision. In: *Proceedings of the 50th Annual Meeting of ASIS;* 24*: 11-13;* Medford (NJ); 1987.

24. Glunz D, Wakiji E. Maximizing search quality through a program of peer review. Online 1983; 7(5):100-110.

25. Hawking DT, Wagers R. Online bibliographic search strategy development. Online 1982; 6(3):12-19.

26. Khan KS, Kunz R, Kleijnen J, Antes G. Systematic reviews to support evidence-based medicine: How to review and apply findings of healthcare research. London: Royal Society of Medicine Press Ltd., 2003.

27. Shiri AA, Revie C, Chowdhury G. Thesaurus assisted search term selection and query expansion: a review of user centred studies. Knowl Organ 2002; 29(1):1-19.

28. Higgins JPT, Green S, editors. *Cochrane Handbook for Systematic Reviews of Interventions 4.2.6* [updated September 2006]. The Cochrane Library, Issue 4, 2006. Chichester (UK): John Wiley & Sons, Ltd; 2006. Available: http://www.cochrane.org/resources/handbook/Handbook4.2.6Sep2006.pdf (accessed 6th October 2006).

29. Drabenstott KM, Weller MS. Handling spelling errors in online catalog searches. Libr Resour Tech Serv 1996; 40(2):113-132.

30. Wildemuth BM, Moore ME. End-user search behaviors and their relationship to search effectiveness. Bull Med Libr Assoc 1995; 83(3):294-304.

31. Lucas W, Topi H. Form and function: the impact of query term and operator usage on Web search results. J Am Soc Inf Sci 2002; 53(2):95-108.

32. Graham RY. Subject no-hits searches in an academic library online catalog: an exploration of two potential ameliorations. Coll Res Libr 2004; 65(1):36-54.

33. Fielding AM, Powell A. Using Medline to achieve an evidence-based approach to diagnostic clinical biochemistry. Ann Clin Biochem 2002; 39(Pt 4):345-350.

34. Barroso J, Gollop CJ, Sandelowski M, Meynell J, Pearce PF, Collins LJ. The challenges of searching for and retrieving qualitative studies. West J Nurs Res 2003; 25(2):153-178.

35. Howard H. Measures that discriminate among online searchers with different training and experience. On-line review 1982; 6(4):315-327.

36. Bysouth PT. Evaluating the use of several approaches to online literature retrieval by research scientists. In: Bysouth PT, editor. End-user Searching: The effective gateway to published Information. London: Aslib, 1990: 105-123.

37. Sewell W, Teitelbaum S. Observations of end-user online searching behavior over eleven years. J Am Soc Inf Sci 1986; 37(4):234-245.

38. Crumley E, Bhatnagar N, Stobart K. Peer reviewing comprehensive search strategies in hemophilia and von Willebrand disease. J Can Health Lib Assoc 2004; 25(4):113-116.

39. Jenuwine ES, Floyd JA. Comparison of Medical Subject Headings and text-word searches in MEDLINE to retrieve studies on sleep in healthy individuals. J Med Libr Assoc 2004; 92(3):349-353.

40. McJunkin MC. Precision and recall in title keyword searches. Inf Tech Lib 1995; 14(3):161-171.

41. Kristensen J. Expanding end-users' query statements for free text searching with a search-aid thesaurus. Inf Process Manag 1993; 29(6):733-744.

42. Keen EM. Some aspects of proximity searching in text retrieval systems. J Info Sci 1992; 18(2):89-98.

43. Keen EM. The use of term position devices in ranked output experiments. J Doc 1991; 47(1):1-22.

44. Lucas W, Topi H. Training for web search: will it get you in shape? J Am Soc Inf Sci Technol 2004; 55(13):1183-1198.

45. Deacon P, Smith JB, Tow S. Using metadata to create navigation paths in the HealthInsite Internet gateway. Health Info Libr J 2001; 18(1):20-29.

46. McKibbon KA, Haynes RB, Dilks CJ, Ramsden MF, Ryan NC, Baker L et al. How good are clinical MEDLINE searches? A comparative study of clinical end-user and librarian searches. Comput Biomed Res 1990; 23(6):583-593.

47. Kim CS. Predicting information searching performance with measures of cognitive diversity[Thesis]. Madison WI : University of Wisconsin, 2002.

48. Haynes RB, Wilczynski N, McKibbon KA, Walker CJ, Sinclair JC. Developing optimal search strategies for detecting clinically sound studies in MEDLINE. J Am Med Inform Assoc 1994; 1(6):447-458.

49. Tenopir C, Shu ME. Magazines in full text: uses and search strategies. Online Rev 1989; 13(2):107-118.

50. Martinez C, Zarember I. OR NOT: the unused operator. J Am Soc Inf Sci 1978; 29(4):207-208.

51. Marlborough HS. Accessing the literature: using bibliographic databases to find journal articles. Part 1. Prim Dent care 2001; 8(3):117-121.

52. Monoi s, O'Hanlon N, Diaz KR. Online searching skills: development of an inventory to assess self-efficacy. J Academic Librarianship 2005; 31(2):98-105.

53. Anonymous. Conducting comprehensive biomedical searches. Biomedical information seminar. The Dialog Corporation, 2003.

54. Critical Appraisal Skills Programme: finding the evidence workshop (CASfew). A 10 step strategy for more effective Medline searches [workshop]. CASPfew; 1998; Oxford (UK).

55. Goodman C. Literature searching and evidence interpretation for assessing health care practices. Stockholm, Sweden: Swedish Council on Technology Assessment in Health Care, 1993:16-32.

56. McKinin EJ, Sievert M, Johnson ED. Using repetition to increase precision in files with large blocks of text. Online Rev 1989; 13:369-382.

57. Jansen BJ, Spink A, Saracevic T. Real life, real users, and real needs: a study and analysis of user queries on the Web. Inf Process Manag 2000; 36(2):207-227.

58. King NS. Search characteristics and the effects of experience on end users of PaperChase. Coll Res Libr 1991; 52(4):360-374.

59. Watson RJ, Richardson PH. Accessing the literature on outcome studies in group psychotherapy: the sensitivity and precision of Medline and PsycINFO bibliographic database searching. Br J Med Psychol 1999; 72(Pt 1):127-134.

60. Anonymous. Development of an optimal search strategy for the retrieval of controlled trials using PubMed. Abstr Workshops Sci Sess Int Cochrane Coll 1998; 6(85).

61. Bickley S. Searching for trials - bridging the gaps. 12th Cochrane Colloquium: Bridging the Gaps; 2004 Oct 2-6; Ottawa.

62. Blackhall K, Crumley E. Setting up search strategies for systematic reviews (or, how many ways can you spell diarrhea?). Bib Medica Can 2003; 24(4):167-168.

63. Evans D. Database searches for qualitative research. J Med Libr Assoc 2002; 90(3):290-293.

64. Galandi D, Bassler D, Antes G. Identifying randomized controlled trials published in German general health care journals using Medline and Embase: how useful is the controlled vocabulary? 8th Annual Cochrane Colloquium; 2000 Oct 25-29; Cape Town (ZA).

65. Guessard B. [Comparison of three bibliographic data bases]. [French]. Soins Form Pedagog Encadr 1999;(30):42-46.

66. Jenkins M. Evaluation of methodological search filters: a review. Health Info Lib J 2004; 21(3):148-163.

67. McCain KW, White HD, Griffith BC. Comparing retrieval performance in online data bases. Inf Process Manag 1987; 23(6):539-553.

68. McKibbon A, Eady A, Marks S. PDQ Evidence-Based Principles and Practice. Hamilton, ON: B.C.Decker Inc., 1999.

69. Watson RJ, Richardson PH. Identifying randomized controlled trials of cognitive therapy for depression: comparing the efficiency of Embase, Medline and PsycINFO bibliographic databases. Br J Med Psychol 1999; 72(Pt 4):535-542.

70. Clarke M., Greaves L., James S. MeSH terms must be used in Medline searches. BMJ 1997; 314:1203.

71. Hutchinson T. Strategies for searching online finding aids: a retrieval experiment. Archivaria 1997;(44):72-101.

72. Burdick AJ. Using unbound subheadings to increase recall in MEDLINE. Bull Med Libr Assoc 1983; 71(3):282-286.

73. Harrison J. Designing a search strategy to identify and retrieve articles on evidence-based health care using MEDLINE. Health Libr Rev 1997; 14(1):33-42.

74. Balas EA, Stockham MG, Mitchell JA, Sievert ME, Ewigman BG, Boren SA. In search of controlled evidence for health care quality improvement. J Med Syst 1997; 21(1):21-32.

75. Plovnick RM, Zeng QT. Reformulation of consumer health queries with professional terminology: a pilot study. J Med Internet Res 2004; 6(3):e27.

76. Barnum C, Henderson E, Hood A, Jordan R. Index versus full-text search: a usability study of user preference and performance. Tech Comm 2004; 51(2):185-206.

77. Muddamalle MR. Natural language versus controlled vocabulary in information retrieval: a case study in soil mechanics. J Am Soc Inf Sci 1998; 49(10):881-887.

78. Spink A, Saracevic T. Interaction in information retrieval: selection and effectiveness of search terms. J Am Soc Inf Sci 1997; 48(8):741-761.

79. Srinivasan P. Optimal document-indexing vocabulary for MEDLINE. Inf Process Manag 1996; 32(5):503-514.

80. Hersh WR, Hickam DH. An evaluation of interactive Boolean and natural language searching with an online medical textbook. J Am Med Inform Assoc 1995; 46(7):478-489.

81. Dickersin K, Scherer R, Lefebvre C. Identifying relevant studies for systematic reviews. BMJ 1994; 309(6964):1286-1291.

82. Richwine PW. A study of MeSH and UMLS for subject searching in an online catalog. Bull Med Libr Assoc 1993; 81(2):229-233.

83. Weinberg BH, Cunningham JA. The relationship between term specificity in MeSH and online postings in MEDLINE. Bull Med Libr Assoc 1985; 73(4):-372.

84. Sievert ME, Boyce BR. Hedge trimming and the resurrection of the controlled vocabulary in online searching. Online Rev 1983; 7(6):489-494.

85. Markey K, Atherton P, Newton C. An analysis of controlled vocabulary and free text search statements in online searches. Online Rev 1980; 4(3):224-236.

86. Savoy J. Bibliographic database access using free-text and controlled vocabulary: an evaluation. Inf Process Manag 2005; 41(4):873-890.

87. 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(1):150-153.

88. Ingui BJ, Rogers MA. Searching for clinical prediction rules in MEDLINE. J Am Med Inform Assoc 2001; 8(4):391-397.

89. Deville WL, Bezemer PD, Bouter LM. Publications on diagnostic test evaluation in family medicine journals: an optimal search strategy. J Clin Epidemiol 2000; 53(1):65-69.

90. van der Weijden T, IJzermans CJ, Dinant GJ, van Duijn NP, de Vet R, Buntinx F. Identifying relevant diagnostic studies in MEDLINE. The diagnostic value of the erythrocyte sedimentation rate (ESR) and dipstick as an example. Fam Pract 1997; 14(3):204-208.

91. 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(3):741-748.

92. Wright LC, Sutherland HJ, Jackson JI, Till JE. Comparison of search strategies on CD Plus/MEDLINE. CMAJ 1991; 145(5):457-464.

93. Fraser C, Thompson MA. Identifying non-randomised studies in Medline. Ann Meeting Int Soc Tech Assess Health Care 1998; 15:113.

94. Crumley E. Search Strategies to Identify Randomized Trials in MEDLINE: A systematic review. Evidence Based Librarianship Conference; 2003 June 4-6; Edmonton.

95. Federiuk CS. The effect of abbreviations on MEDLINE searching. Acad Emerg Med 1999; 6(4):292-296.

96. Wieland S, Brodney S, Dickersin K. Designing an efficient and precise search strategy for observational studies. 10th Cochrane Colloquium; 2002 Jul 31-Aug 3; Stavanger (NO).

97. Wilczynski NL, Walker CJ, McKibbon KA, Haynes RB. Quantitative comparison of pre-explosions and subheadings with methodologic search terms in MEDLINE. Proc Annu Symp Comput Appl Med Care 1994;905-909.

98. Jadad AR, McQuay HJ. Searching the literature : be systematic in your searching. BMJ 1993; 307(6895):66.

99. Gotzsche PC, Lange B. Comparison of search strategies for recalling double-blind trials from MEDLINE. Dan Med Bull 1991; 38(6):476-478.

100. Funk ME, Reid CA. Indexing consistency in MEDLINE. Bull Med Libr Assoc 1983; 71(2):176-183.

101. Booth A. "Brimful of STARLITE": Towards standards for reporting literature searches. HTA: Health Technology Assessment International 1st Annual Meeting; 2004 May 31-Jun 2;.Krakow (PL).

102. Coletti MH, Bleich HL. Medical Subject Headings Used to Search the Biomedical Literature. J Am Med Inform Assoc 2001; 8(4):317-323.

103. Lowe HJ, Barnett GO. Understanding and using the medical subject headings (MeSH) vocabulary to perform literature searches. JAMA 1994; 271(14):1103-1108.

104. Leininger K. Interindexer consistency in PsycINFO. J Libr Info Sci 2000; 32(1):4-8.

105. Zacks MP, Hersh WR. Developing search strategies for detecting high quality reviews in a hypertext test collection. Proc AMIA Symp 1998;663-667.

106. Rowley J. The controlled versus natural indexing languages debate revisited: a perspective on information retrieval practice and research. J Info Sci 1994; 20(2):108-119.

107. Mijnhout GS, Hooft L, van Tulder MW, Deville WL, Teule GJ, Hoekstra OS. How to perform a comprehensive search for FDG-PET literature. Eur J Nucl Med 2000; 27(1):91-97.

108. Othman R. An applied ethnographic method for evaluating retrieval features. Electronic Lib 2004; 22(5):425-432.

109. Lobo DO. Metodos y tecnicas para la indizacion y recuperacion de los recursos de la World Wide Web. Methods and techniques for indexing and retrieving World Wide Web resources. Boletin de la Asociacion Andaluza de Bibliotecarios 1999; 14(57):11-22.

110. Jensen MF, Ket H. Check it out! A checklist for evaluating the reporting of literature search methodology in HTAs and CPGs [poster]. DACEHTA 6th Symposium HTA; 2005 Nov 3-4; Cologne (DE).

111. Moher D, Cook DJ, Eastwood S, Olkin I, Rennie D, Stroup DF. Improving the quality of reports of meta-analyses of randomised controlled trials: the QUOROM statement. Br J Surg 2000; 87(11):1448-1454.

112. Felson DT. Bias in meta-analytic research. J Clin Epidemiol 1992; 45(8):885-892.

113. Egger M, Smith GD. Bias in location and selection of studies. BMJ 1998; 316(7124):61-66.

114. McGowan J. Literature searching in Evidence-based Rheumatology. In: Tugwell P, Shea B, editors. *Evidence-based Rheumatology*. London 2005. p.3-18.

115. Ludl H, Schope K, Mangelsdorf I. Searching for information on chemical substances in selected biomedical bibliographic databases. Chemosphere 1995; 31(2):2611-2628.

116. Buntrock RE. Effect of search environment on search performance. Online 1979; 3(4):10-13.

117. Wilczynski NL, McKibbon KA, Haynes RB. Response to Glanville et al.: How to identify randomized controlled trials in MEDLINE: ten years on. J Med Libr Assoc 2007; 95(2):117-118.

118. Jenkins M. Evaluation of methodological search filters--a review. Health Info Libr J 2004; 21(3):148-163.

119. Glanville JM, Lefebvre C, Miles JN, Camosso-Stefinovic J. How to identify randomized controlled trials in MEDLINE: ten years on. J Med Libr Assoc 2006; 94(2):130-136.

120. Dickersin K, Scherer R, Lefebvre C. Identifying relevant studies for systematic reviews. BMJ 1994; 309(6964):1286-1291.

121. Robinson KA, Hinegardner PG, Lansing P. Development of an optimal search strategy for the retrieval of controlled trials using PubMed. 6th Annual Cochrane Colloquium; 1998 Oct 22-26; Baltimore.

122. Moher D, Tetzlaff J, Tricco AC, Sampson M, Altman DG. Epidemiology and reporting characteristics of systematic reviews. PLoS Med 2007; 4(3):e78.

123. Sampson M, McGowan J, Lefebvre C, Grimshaw JM, Moher D. PRESS (Peer Review Electronic Search Strategy): a proposal to develop a quality assessment checklist and expert peer review forum for HTA searches. CADTH Final Performance Report. Ottawa: Canadian Agency for Drugs and Technologies in Health; 2006.

124. Egger M, Juni P, Bartlett C, Holenstein F, Sterne JA. How important are comprehensive literature searches and the assessment of trial quality in systematic reviews? Empirical study. Health Technol Assess 2003; 7(1):1-76.

125. Armour T, Dingwall O, Sampson M. Contribution of checking reference lists to systematic reviews [poster]. XIII Cochrane Colloquium; 2005 Oct 22; Melbourne.

126. Walker G. *The search performance of end-users. Proceedings of the Ninth National Online Meeting; 1988 May 10-12; New York*. Medford (NJ), Learned Information Inc.; 1988. 403-410 s 1988;-12.

127. Moher D, Pham B, Lawson ML, Klassen TP. The inclusion of reports of randomised trials published in languages other than English in systematic reviews. Health Technol Assess 2003; 7(41):1-90.

128. Egger M, Zellweger-Zahner T, Schneider M, Junker C, Lengeler C, Antes G. Language bias in randomised controlled trials published in English and German. Lancet 1997; 350:326-329.

129. Iansavichene AI, Sampson M, McGowan J, Ajiferuke I. Should systematic reviewers search for randomized controlled trials published as letters? Canadian Health Libraries Association Conference; 2007 May 28-June 01; Ottawa.

130. Higgins JPT and Green S. Cochrane Handbook for Systematic Reviews of Interventions 4.2.5 [web site]. John Wiley & Sons, Ltd. Chichester (UK): John Wiley & Sons, Ltd; 2005. Available: http://www.cochrane.org/resources/handbook/hbook.htm

131. McGowan J, Sampson M, Santesso N. Collection development in support of Cochrane reviews: Normative data on sources and database interface from 105 Cochrane reviews [poster]. UNYOC Annual Conference; 2004 Oct 13; Ottawa.

APPENDICES

Available from CADTH's web site
www.cadth.ca