McMaster University – HiRU’s Approach to Search Filter Development

Author: Nancy Wilczynski PhD

Date: 2011-09-27

Health Information Research Unit
Evidence-Based Health Informatics

HiRU
HEALTH INFORMATION
RESEARCH UNIT

This handout describes an approach to search filter development and validation that has been very successful. Researchers in the Health Information Research Unit (HiRU) at McMaster University have a way for identifying candidate search terms that works extremely well. This approach will be described. We also have a database (~50,000 tagged articles; includes both on-target and off-target articles) that can be used by others who want to try out their own "tools". Examples of our collaborative efforts with other researchers who have used this database for machine learning purposes are provided. Additionally, we describe a "tool" we developed in-house for identifying frequently occuring terms that can be NOTed out to increase search precision.

HiRU‘s Approach to Search Filter Development [1-3]

The search filters developed and validated by researchers in HiRU are available for use on PubMed’s Clinical Queries page (http://www.ncbi.nlm.nih.gov/pubmed/clinical) and Health Services Research Queries page (http://www.nlm.nih.gov/nichsr/hedges/search.html); Ovid’s Additional Limits page for MEDLINE, Embase and PsycINFO; EBSCOhost’s main searching page for CINAHL; and HiRU’s Nephrology Filters page (http://hiru.mcmaster.ca/hiru/HIRU_Hedges_Nephrology_Filters.aspx). The study design used to develop these filters and others was an analytic survey using a diagnostic testing framework. For each of the filters developed, research staff reviewed a pre-defined set of journals across various publishing years and categorized all articles according to pre-specified criteria. A list of search terms were compiled after seeking input from individuals and groups including clinicians, expert searchers, librarians, and the US National Library of Medicine (NLM). The proposed search filters were treated as “diagnostic tests”, and the manual review (hand search) of the literature was treated as the “gold standard”. We determined the sensitivity, specificity, precision, and accuracy of each single term and combinations of terms using an automated process developed in-house. An overview of the methods used to develop the methodologic filters that are available for use on

PubMed’s Clinical Queries page has been published (http://www.biomedcentral.com/1472-6947/5/20)[1].

The details for identifying candidate search terms for use when developing the methodologic search filters (e.g., filters developed to detect methodologically rigourous studies about treatment) follows (a similar approach was taken when developing all search filters):

We began a list of index terms and textwords for each of the four electronic databases, MEDLINE, Embase, CINAHL, and PsycINFO, and then sought input from clinicians and librarians in the United States and Canada through interviews of known searchers, requests at meetings and conferences, and requests to the US NLM. Individuals were asked what terms or phrases they used when searching for studies of causation, prognosis, diagnosis, treatment, economics, clinical prediction guides, reviews, costs, and of a qualitative nature when using these databases. For instance, for MEDLINE, terms could be from Medical Subject Headings (MeSH), including publication types (pt), and subheadings (sh), or could be textwords (tw) denoting methodology in titles and abstracts of articles. The number of terms compiled for each of the electronic databases is shown in the table below. Index terms varied by electronic database whereas the same list of textwords were tested in each of the electronic databases. The list of search terms used for testing in each of the electronic database are availabe upon request.

Table. Number of Terms Compiled for Each of the Electronic Databases

Database	No. of Terms	No. of Unique Terms	No. of Terms that Returned Results
MEDLINE	5345	4862	3870
Embase	5385	4843	3542
CINAHL	5020	5020	3110
PsycINFO	4985	4985	2583

After the list of terms was compiled and the manual review of the literature was completed, the following provides an example of the approach taken when developing the methodologic search filters for detecting diagnostic studies in MEDLINE (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC403841/?tool=pubmed)[2]:

Individual search terms with a sensitivity of more than 25% and a specificity of more than 75% for the diagnosis category were incorporated into the development of search filters that included a combination of two or more terms. All combination of terms used the Boolean OR – for example, “sensitivity.tw. OR specificity.tw.” (Ovid syntax). For the development of multiple term search filters to optimize either sensitivity or specificity, we tested the combination of individual terms with all two term search filters with sensitivity at least 75% and specificity at least 50%. For optimizing accuracy, two term search filters with accuracy of more than 75% were considered for multiple term development. Overall, we tested 17,287 multiple term search filters. Search filters were also developed that optimized combined sensitivity and specificity (equivalent to the optimal point on a receiver operating characteristic curve, minimizing the total number of errors).

Our approach to search filter development (i.e., compiling a list of search terms using non-automated means and testing them using software developed in-house) has been very successful. Sensitivities and specificities as high as 99.9% have been achieved (see HiRU Hedges website for some examples, http://hiru.mcmaster.ca/hiru/HIRU_Hedges_home.aspx), and search filter performance is better than or comparable to filters developed using other means [3].

Clinical Hedges Database Available for Use by Others

Our Clinical Hedges Database which contains data compiled from the manual review of the literature is available for use by others who want to try out their own "tools". Our MEDLINE Clinical Hedges Database contains 49,028 unique records of articles published in 161 journals in the year 2000. All records are indexed in MEDLINE and have been categorized by highly calibrated research staff for numerous features including the following:

Format (categories and definitions)

Original study - any full text article in which the investigators report first-hand observations.

Review - any full text article that is bannered ‘review, overview, or meta-analysis’ in the title or in a section heading, or it is indicated in the text of the article that the intention was to review, summarize, highlight, etc. the literature on a particular topic.

General and miscellaneous articles - a general or philosophical discussion of a topic without original observation and without a statement that the purpose was to review or appraise a body of knowledge. This could include news items, unbannered editorials, bannered and unbannered conference reports, position and opinion papers, musings, psychosocial observations, and decision analysis that cannot be classified as an original study or review.

Case report - is an original study or report that presents only individualized data. The data are not combined in any way, and often involves less than 10 subjects.

Of interest to human health care

Yes = concerned with the understanding of health care in humans; anything that will have an effect on the patient/subject (e.g., promoting wellness).

No = not concerned with the understanding of health care in humans; anything that will not have an effect on the patient/subject (e.g., studies that describe the normal development of people; basic science; studies involving animals, gender and equality studies in the health profession, the medical profession looking at itself, studies looking only at the structure and organization of the health care system, or studies looking at research methodology issues).

Purpose (categories and definitions)

Etiology - Content pertains directly to determining if there is an association (causal link) between an exposure and **a disease or condition** (examples of a condition are low birth weight, large [or small] for gestational age, preterm birth, miscarriage, abortion, cesarean section, pregnancy, or death). The question that is being asked is **“What causes people to get a disease or condition?”**

Prognosis - Content pertains directly to the prediction of the clinical course or the natural history of a **disease or condition** (examples of a condition are low birth weight, large [or small] for gestational age, preterm birth, or pregnancy) **with the disease or condition existing at the beginning of the study**.

Diagnosis - Content pertains directly to using a tool to arrive at a diagnosis of **a disease or condition**. **Screening** to make a diagnosis is included here.

Treatment, prevention, continuing medical education or quality improvement - Content pertains directly to therapy (including adverse effects studies), prevention, rehabilitation, quality improvement, or continuing medical education. For a study to be classified as therapy (which includes prevention, continuing medical education and quality improvement) the investigators must intervene – there has to be an intervention that can be manipulated.

Costs -Content pertains directly to the costs or financing or economics of a health care issue.

Economics - Content pertains directly to the economics of a health care issue. The economic question addressed must be based on **comparison of alternatives**, i.e., comparison of the costs and effects of at least 2 different forms of intervention or service provision. Thus, 'costing' or 'financing' of a single health service, even if for a variety of conditions, does not constitute an economic study; an economic study would compare 2 (or more) different ways of providing the same service, and would include at least intermediate (e.g., BP) or more advanced (e.g., stroke) outcomes.

Clinical prediction guide - Content pertains directly to the prediction of some aspect of a disease or condition; the authors must indicate that the purpose of the study is **to develop or validate a rule, guide, index, equation, scale, score or model to predict** a diagnosis, prognosis, risk (etiology), therapeutic response, therapeutic drug levels or clinical outcome. For everything except diagnosis the patients must be followed over time.

Qualitative study - Content of study contains the following qualities: The content relates to how people feel or experience certain situations, specifically those situations that relate to health care in humans; Collection methods are appropriate for qualitative data; Analyses are appropriate for qualitative data.

Something else - The purpose of the study does not fit any of the above.

Rigour (categories and methodologic criteria)

Etiology - Observations concerned with the relationship between exposures and putative clinical outcomes; Data collection is prospective; Clearly identified comparison group(s); Blinding of observers of outcome to exposure.

Prognosis - Inception cohort of individuals all initially free of the outcome of interest; Follow-up of ≥80% of patients until the occurrence of a major study end point or to the end of the study; Analysis consistent with study design.

Diagnosis - Inclusion of a spectrum of participants; Objective diagnostic ("gold") standard OR current clinical standard for diagnosis; Participants received both the new test and some form of the diagnostic standard; Interpretation of diagnostic standard without knowledge of test result and vice versa; Analysis consistent with study design.

Treatment or Prevention (including quality improvement and continuing medical education studies) - Random allocation of participants to comparison groups; Outcome assessment of at least 80% of those entering the investigation accounted for in 1 major analysis at any given follow up assessment; Analysis consistent with study design.

Economic studies - Question is a comparison of alternatives; Alternative services or activities compared on outcomes produced (effectiveness) and resources consumed (costs); Evidence of effectiveness must be from a study of real patients that meets the above-noted criteria for diagnosis, treatment, quality improvement, or a systematic review article; Effectiveness and cost estimates based on individual patient data (micro-economics); Results presented in terms of the incremental or additional costs and outcomes of one intervention over another; Sensitivity analysis if there is uncertainty**.**

Clinical prediction guides - Guide is generated in one or more sets of real patients (training set); Guide is validated in another set of real patients (test set).

Reviews - Statement of the clinical topic; Explicit statement of the inclusion and exclusion criteria; Description of the methods; ≥ 1 article must meet the above noted criteria.

Examples of Collaborations with Other Researchers Who Have Used Our Clinical Hedges Database [4]

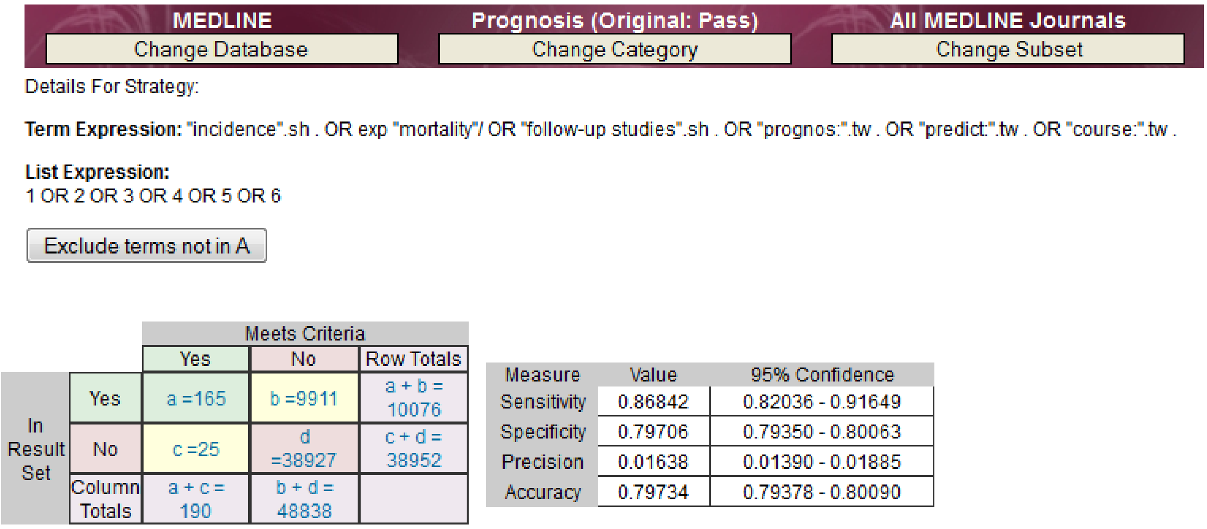
Other methods have been proposed and studied for developing search filters. We collaborated with Kilicoglu and colleagues at the US NLM who approached the problem of recognizing studies containing useable clinical advice from retrieved topically relevant articles as a binary classification problem http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2605595/?tool=pubmed[4]. The gold standard used in the development of PubMed clinical query filters (HiRU's Clinical Hedges Database) formed the basis of their approach. They identified scientifically rigorous studies using supervised machine learning techniques (Naïve Bayes, support vector machine [SVM], and boosting) trained on high-level semantic features. They combined these methods using an ensemble learning method (stacking). The performance of learning methods was evaluated using precision, recall and F_1 score, in addition to area under the receiver operating characteristic (ROC) curve (AUC). Using a training set of 10,000 manually annotated MEDLINE citations (from HiRU's Clinical Hedges Database), and a test set of an additional 2,000 citations (from HiRU's Clinical Hedges Database), they achieved 73.7% precision and 61.5% recall in identifying rigourous, clinically relevant studies, with stacking over five feature-classifier combinations and 82.5% precision and 84.3% recall in recognizing rigourous studies with treatment focus using stacking over *word + metadata* feature vector. Their results demonstrated that a high quality gold standard and advanced classification methods can help clinicians acquire best evidence from the medical literature.

We also collaborated with Zhang and colleagues at the University of Wisonsin-Milwaukee who used HiRU's Clinical Hedges Database to develop superivsed machine-learning techniques to automatically classify an article into four clinically useful distinquishing formats: original study, review, case report and general. They also made a distinction between articles of interest to human health care (HHC) and those that were not. Futhermore, for those articles that fell into the original and review categories, as well as HHC, they further classified them into five categories of purpose: etiology, diagnosis, treatment, prognosis, and others. They explored different machine-learning models, including Support Vector Machines, Naïve Bayes Multinomial, AdaBoost.M1, and Stacking. They currently are in the process of publishing their findings.

We provided the above-mentioned reseachers with data from the Clinical Hedges Database. We are willing to collaboration with others (see contact information below).

Tool Developed in HiRU for Identifying Frequently Occurring Terms

The programmers in HiRU developed a tool to determine the term frequency to answer the following research question: Can search filter precision be improved by NOTing out the text words and index terms assigned to those articles that are retrieved but are off-target? The tool was developed for use in a web-based interface that allows the user to built search filters for a specified electronic database (e.g., MEDLINE), purpose category (e.g., prognosis) and journal subset (e.g., all MEDLINE journals). After a search filter is built the following information is presented to the user:



MEDLINE | Prognosis (Original: Pass) | All MEDLINE Journals

Details For Strategy:

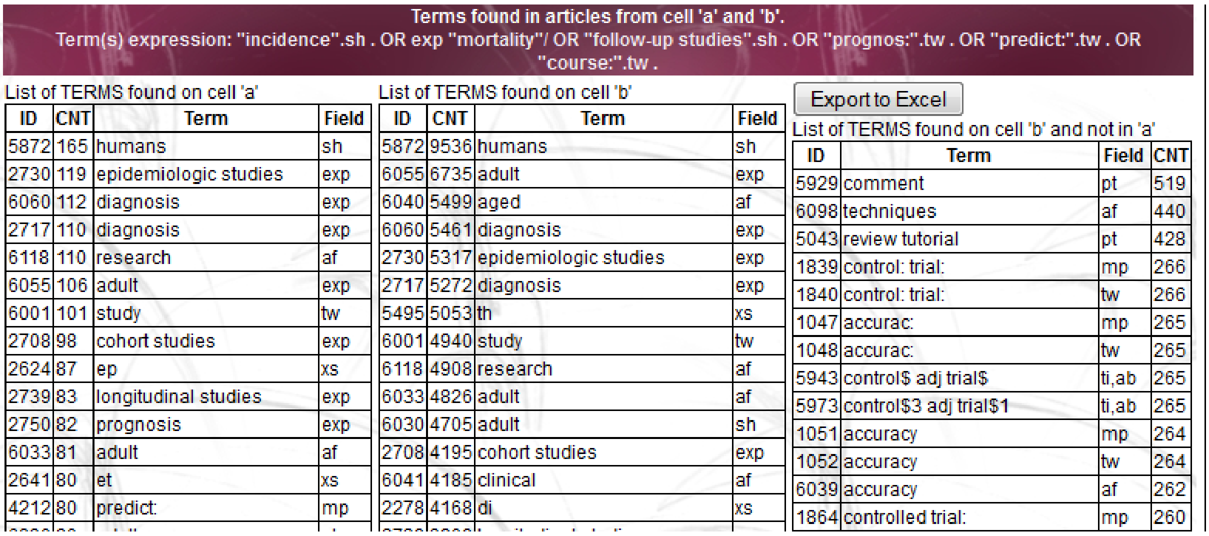
Term Expression: "incidence".sh . OR exp "mortality"/ OR "follow-up studies".sh . OR "prognos:".tw . OR "predict:".tw . OR "course:".tw .

List Expression:
1 OR 2 OR 3 OR 4 OR 5 OR 6

		Meets Criteria		
		Yes	No	Row Totals
In Result Set	Yes	a =165	b =9911	a + b = 10076
	No	c =25	d =38927	c + d = 38952
	Column Totals	a + c = 190	b + d = 48838	

Measure	Value	95% Confidence
Sensitivity	0.86842	0.82036 - 0.91649
Specificity	0.79706	0.79350 - 0.80063
Precision	0.01638	0.01390 - 0.01885
Accuracy	0.79734	0.79378 - 0.80090

As shown above when determining the operating characteristic of search filters in the Clinical Hedges interface each cell contains the number of ciations retrieved. Within the Clinical Hedges web interface the user can retrieve the citations noted in cells A (165 citations in cell A – shown above) and B (9,911 citations in cell B – shown above). Using a program developed in-house we created a list and rank ordered all text words and index terms (from the list of terms tested when developing the search filters) that were unique to citations found in cell B when compared with citations found in cell A. The program that was developed in-house is similar to Microsoft Office Index function. Cell B citation terms needed to be unique when compared with cell A citations to ensure that sensitivity did not decrease when NOTing out content. Output from the program is shown below:



Terms found in articles from cell 'a' and 'b'.
Term(s) expression: "incidence".sh . OR exp "mortality"/ OR "follow-up studies".sh . OR "prognos:".tw . OR "predict:".tw . OR "course:".tw .

List of TERMS found on cell 'a'

ID	CNT	Term	Field
5872	165	humans	sh
2730	119	epidemiologic studies	exp
6060	112	diagnosis	exp
2717	110	diagnosis	exp
6118	110	research	af
6055	106	adult	exp
6001	101	study	tw
2708	98	cohort studies	exp
2624	87	ep	xs
2739	83	longitudinal studies	exp
2750	82	prognosis	exp
6033	81	adult	af
2641	80	et	xs
4212	80	predict:	mp

List of TERMS found on cell 'b'

ID	CNT	Term	Field
5872	9536	humans	sh
6055	6735	adult	exp
6040	5499	aged	af
6060	5461	diagnosis	exp
2730	5317	epidemiologic studies	exp
2717	5272	diagnosis	exp
5495	5053	th	xs
6001	4940	study	tw
6118	4908	research	af
6033	4826	adult	af
6030	4705	adult	sh
2708	4195	cohort studies	exp
6041	4185	clinical	af
2278	4168	di	xs

List of TERMS found on cell 'b' and not in 'a'

ID	Term	Field	CNT
5929	comment	pt	519
6098	techniques	af	440
5043	review tutorial	pt	428
1839	control: trial:	mp	266
1840	control: trial:	tw	266
1047	accurac:	mp	265
1048	accurac:	tw	265
5943	control$ adj trial$	ti,ab	265
5973	control$3 adj trial$1	ti,ab	265
1051	accuracy	mp	264
1052	accuracy	tw	264
6039	accuracy	af	262
1864	controlled trial:	mp	260

The unique text words or index terms (shown above in the table labelled List of Terms found on cell b but not in a) were appended as a string of ORed terms using the Boolean operator NOT. For example, (incidence.sh. OR exp mortality OR follow-up studies.sh. OR prognos.tw. OR predict.tw. OR course.tw.) NOT (comment.pt OR techniques.af. OR review tutorial.pt.). The process of developing the ORed string of terms to be NOTed out was automated using software developed in-house. All terms unique to cell B citations were considered starting with the term assigned to the most citations. A term was included in the ORed string if the number of citations in cell B decreased when the term was added.

The results of this research are being presented at the American Medical Informatics Association (AMIA) Annual Symposium 2011 in Washington, DC, October 22-26.

Contact Information

Nancy Wilczynski PhD, wilczyn@mcmaster.ca, 905-525-9140 ext. 22780.

References

1. Wilczynski NL, Morgan D, Haynes RB and the Hedges Team. An overview of the design and methods for retrieving high-quality studies for clinical care. BMC Med Inform Decision Making. 2005 Jun 21;**5:**20. http://www.biomedcentral.com/1472-6947/5/20

2. Haynes RB, Wilczynski NL. Optimal search strategies for retrieving scientifically strong studies of diagnosis from MEDLINE: an analytic survey. BMJ. 2004;328:1040–2. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC403841/?tool=pubmed

3. McKibbon KA, Wilczynski NL, Haynes RB; Hedges Team. Retrieving randomized controlled trials from medline: a comparison of 38 published search filters. Health Info Libr J. 2009 Sep;26(3):187-202. PubMed PMID: 19712211.

4. Kilicoglu H, Demner-Fushman D, Rindflesch TC, **Wilczynski NL**, Haynes RB. Towards automatic recognition of scientifically rigorous clinical research evidence. J Am Med Inform Assoc. 2009 Jan-Feb;16(1):25-31. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2605595/?tool=pubmed