Identifying diagnostic accuracy studies in EMBASE

By Lucas M. Bachmann, M.D.
lucas.Bachmann@evimed.ch
Research Fellow

Horten Centre
Bolleystrasse 40
Postfach Nord
University of Zürich
CH-8091 Zürich
Switzerland

Birmingham University
Birmingham B15 2TG
United Kingdom

Pius Estermann, M.D.
ester@uszbib.unizh.ch
Information Specialist

University Hospital Library
Zurich University
Rämistrasse 100
8091 Zürich
Switzerland

Corinna Kronenberg, M.D.
ac-kronenberg@bluewin.ch
Research Fellow

Horten Centre
Bolleystrasse 40
Postfach Nord
University of Zürich
CH-8091 Zürich
Switzerland

Gerben ter Riet, M.D., Ph.D.
g.terriet@amc.uva.nl
Senior Research Fellow

Horten Centre
Bolleystrasse 40
Postfach Nord
University of Zürich
CH-8091 Zürich
Switzerland

Department of General Practice
Academic Medical Center
Amsterdam Center for Health and Health Care Research (AmCOGG)
Meibergdreef 15
1105 AZ Amsterdam
The Netherlands

Objective: The objective was to develop and test search strategies to identify diagnostic articles recorded on EMBASE.

Methods: Four general medical journals were hand searched for diagnostic accuracy studies published in 1999. Identified studies served as a gold standard. Candidate terms for search strategies were identified using a word-frequency analysis of their abstracts. According to the frequency of identified terms, searches were run for each term independently. Sensitivity, precision, and number needed to read (NNR) (1/precision) of every candidate term were calculated. Terms with the highest ''sensitivity*precision'' product were used as free-text terms and combined into a final strategy using the Boolean operator ''OR.''

Results: The most frequently occurring eight terms (sensitiv* or detect* or accura* or specific* or reliab* or positive or negative or diagnos*) produced a sensitivity of 100% (95% confidence interval [CI] 94.1 to 100%) and an NNR of 27 (95% CI 21.0 to 34.8). The combination of the two truncated terms sensitiv* or detect* gave a sensitivity of 73.8% (95% CI 60.9 to 84.2%) and an NNR of 5.7 (95% CI 4.4 to 7.6).

Conclusions: The identified search terms offer the choice of either reasonably sensitive or precise search strategies for the detection of diagnostic accuracy studies in EMBASE. The terms are useful both for busy health care professionals who value precision and for reviewers who value sensitivity.

INTRODUCTION

When producing systematic reviews, researchers should try to identify as much empirical evidence as possible to inform the review question. Usually, the major biomedical databases such as MEDLINE and EMBASE are the starting points when trying to identify this evidence. However, information retrieval in such databases can become very time consuming because searches usually identify many irrelevant articles (low retrieval precision). In recent years, researchers have adopted various approaches to the development of search strategies to identify different types of studies (therapy, prognosis, diagnosis, and etiology) and different study designs [1–5]. Search strategies to identify diagnostic studies have also been developed [6–9].

For example, in 1994, Haynes and coworkers published a MEDLINE search filter for diagnosis [10], which is now publicly available in PubMed (Clinical Queries) [11]. However, differences in indexing hampered the straightforward use of this filter in EMBASE [12]. For example, the suggested MEDLINE Medical Subject Headings (MeSH) term ''Sensitivity and Specificity'' was only entered into the EMBASE EMTREE thesaurus in 2001. Alternatively, EMTREE provides the controlled vocabulary term ''diagnostic accuracy.'' The authors developed and tested search strategies to identify diagnostic articles recorded on EMBASE.

METHODS

One reviewer (Kronenberg) hand searched all issues published in 1999 of the *New England Journal of Medicine, The Lancet, JAMA,* and *British Medical Journal (BMJ).* The journals used in this study are indexed cover to cover in EMBASE. An article was deemed to be about diagnostic accuracy if at least one test was compared with a reference standard. A test was defined as any procedure used to change the estimate of the likelihood of disease presence. This definition included history taking, physical exam, and more advanced tests. All references of diagnostic studies identified (gold standard) were stored in a Reference Manager file.* Articles that were not diagnostic studies were excluded.

The result of the hand search was assumed to be perfect and reflected the true number of diagnostic accuracy studies in our total set or universe. The challenge for any automated search is to find all the references to accuracy studies (100% sensitivity) and at the same time not to find references to any other studies (100% specificity).

To assess the reproducibility of the hand search, a second reviewer (Bachmann) independently duplicated the hand search in a randomly selected 10% of all issues. The 10% sample was determined by numbering all references in the four journals sequentially, and 10% of references were then randomly selected using the Statistix software.†

The gold standard references were identified in EMBASE (Datastar version) using the accession number, a unique identifier for a specific record. A strategy combining all accession numbers using the Boolean connector ''OR'' was saved. Thus, a search in EMBASE would uniquely identify the gold standard references. The number of references in EMBASE was reduced to the subset of all references (6,143) that were published in the four chosen journals in 1999 to proxy a ''universe'' of searchable articles.

* Information about Reference Manager 9.5, used in this study, may be viewed at http://www.refman.com.

† Information about Statistix 7, used in this study, may be viewed at http://www.statistix.com.

Table 1
Example of term frequencies for the letter ''d'' as provided by the List Index function

Frequency	Location in abstract	Term
2	Keyword3	differential
2	Abstract	differential
1	Keyword3	differentiation
3	Abstract	difficult
1	Abstract	diffuse
1	Keyword1	digene
1	Keyword3	dilatation
1	Abstract	dilatation
1	Author	dilaveris
1	Institution	dim
1	Keyword3	dimensional
1	Abstract	dimensional
1	Keyword3	dimer
1	Abstract	dimer
1	Author	dipiro
4	Abstract	direct
1	Abstract	directed
1	Abstract	directive
1	Title	directives
1	Abstract	directives
4	Abstract	directly
1	Title	disability
1	Keyword3	disability
2	Abstract	disability
1	Abstract	discharge
1	Abstract	discomfort
1	Abstract	discrepancy
1	Keyword3	discriminant
1	Abstract	discriminant
1	Abstract	discriminate
1	Abstract	discriminated
1	Abstract	discriminating
1	Abstract	discrimination
7	Title	disease
16	Keyword3	disease
17	Abstract	disease
2	Title	diseases
3	Abstract	diseases
15	SubjectHeading	diseases
1	Keyword3	dislocation
1	Abstract	dislocation
4	Keyword3	disorder
4	Abstract	disorder
2	Title	disorders
3	Abstract	disorders
1	Abstract	display
1	Abstract	distinct
1	Abstract	distinguish
2	Abstract	distinguished
1	Abstract	distribution
1	Abstract	disturbance
1	Institution	div
6	Institution	division
1	Institution	dk
1	Abstract	dl
3	Title	dna
1	RegNo	dna
4	Keyword3	dna
4	Abstract	dna
1	Author	dobbins
1	Author	doble
1	Author	dobs
1	Abstract	doctor
1	Abstract	donation
1	Title	donations
1	Abstract	donations
1	Keyword3	donor
1	Abstract	donor
1	Institution	donor
1	Abstract	donors
1	Title	doping
1	Keyword3	doping
1	Title	doppler
2	Keyword3	doppler
2	Abstract	doppler
1	Institution	dor
1	Author	dore
2	Keyword3	dose
1	Abstract	dose
1	Abstract	double
1	Author	douglas
1	Title	down
1	Keyword3	down
1	Abstract	down
1	Institution	dr
1	Abstract	drawn
1	Author	drewe
1	Abstract	dried
1	Keyword3	drug
1	SubjectHeading	drug
1	Abstract	drugs
1	Abstract	dtp
1	Institution	du
1	Author	duarte
1	Keyword3	duct
1	Abstract	ductal
2	Abstract	due
1	Author	duffy
1	Author	duggan
1	Author	dunn
1	Institution	dunstans
1	Keyword3	dura
1	Title	dural
1	Abstract	dural
2	Abstract	duration
1	Author	durfee
1	Author	durie
1	Title	during
8	Abstract	during
1	Abstract	dvt
1	Title	dysfunction
2	Abstract	dysfunction
1	Keyword3	dyskaryosis
2	Abstract	dyskaryosis
1	Keyword3	dyspepsia
1	Abstract	dyspepsia
1	Keyword3	dyspnea

Because our primary aim was to define a search strategy using the offered thesaurus terms, we ran a first search with the EMTREE thesaurus term ''Diagnostic Accuracy.'' This term was considered an equivalent of the MEDLINE term ''Sensitivity and Specificity'' by the authors. This search identified 94% of the gold standard studies but was also associated with a low precision of 4.2%. Because further terms added to this EMTREE term would increase the sensitivity but would also further decrease precision, this preliminary finding suggested that we explore the effects focusing only on text words.

Realizing that the identification of relevant text words might be subjective and be associated with substantial risk of bias, we decided to apply the method of Boynton and coworkers [13] who selected potentially useful text words through the process of word frequency analysis.

We performed the frequency analysis of the occurrence of each word in each reference using the Idealist bibliographic software package.‡ The ListIndex func-

‡ Information about Blackwell Idealist, used in this study, may be viewed at http://www.blackwell-science.com/Products/IDEALIST/DEFAULT.HTM.

Table 2
List of twenty-three (truncated) terms with corresponding sensitivities and precisions if searched as a single term

Term (truncated)	Sensitivity (%)	Precision (%)
diagnos*	93.44	4.40
detect*	57.38	19.66
test*	52.46	7.80
accura*	50.82	21.38
control*	49.18	2.44
analy*	47.54	2.85
sensitiv*	45.90	27.72
specific	44.26	12.56
measure*	42.62	4.18
screen*	40.98	8.93
high*	40.98	3.65
assess*	39.34	4.17
positive	34.43	13.13
risk	34.43	1.93
interpretation*	31.15	5.94
identif*	27.87	6.25
normal	27.87	5.94
negative	26.23	16.00
predict*	26.23	9.52
examination*	18.03	8.09
determine	18.03	4.82
reliab*	16.39	29.41

tion in the software was used to determine the frequency of occurrence of all the words in the titles, abstracts, and subject index. An example for words starting with the letter ''d'' is provided in Table 1.

The list was transferred to a Microsoft Excel file. To specifically select terms semantically associated with diagnostic accuracy, two reviewers (Estermann and Bachmann) excluded numbers, single letters, author names and institutions, register numbers, and journal names. Terms were also excluded if they were general medical language, for example, organ names or diseases, population of interest, or the word ''study.'' We considered these words not helpful in focusing a search on diagnostic accuracy studies. If the two reviewers disagreed on excluding a term, it was included. All included expressions were sorted alphabetically.

If terms differed only in the ending (e.g., diagnosis, diagnose, diagnostic, diagnostics), we decided to use the truncated term (e.g., ''diagnos*''). According to the frequency of the identified most-frequent (truncated) terms, twenty-three searches were run for each term independently (Table 2). Sensitivity (retrieved articles as a proportion of all gold standard diagnostic articles), precision (gold standard diagnostic articles as a proportion of all retrieved articles), and number needed to read (NNR = 1/precision) of each text word were then calculated. Sensitivity is the number of electronically retrieved citations as a proportion of the number of truly relevant full papers (or diagnostic accuracy studies in this paper). The term is often used in medical research on diagnostic tests where it reflects the proportion of persons with a non-normal test result among all patients with some target disease as established by a gold standard reference test. Precision is the number of relevant full papers as a proportion of the number of electronically retrieved citations. In a clinical context, this quantity is often called ''positive predictive value,'' the proportion of persons with the target disease among persons with a non-normal test result. In addition, we coined the term NNR as an analogy to the number needed to treat (NNT) to describe the number of irrelevant references that have to be screened to find one of relevance. The NNR refers to the number of titles or abstracts necessary to read and ponder to find a reference to another relevant study in the set of retrieved references.

Next, the product of sensitivity and precision was computed for each of the text words. We decided to calculate this figure because we wanted to identify those terms most balanced for sensitivity and precision. We thought that only terms contributing both to sensitivity and to precision were useful in building an efficient strategy.

The ten terms with the highest sensitivity-precision product were combined using ''OR'' to produce a series of search strategies. The sensitivities and precisions of these cumulative search strategies were then calculated.

RESULTS

The hand searches identified sixty-one articles (gold standard) as citations of a diagnostic accuracy study from a pool of 6,143 references in the four selected journals. We assumed these sixty-one citations to be the true number of diagnostic papers in the set of 6,143 references. The twenty-three truncated terms with the highest frequency according to the ListIndex function (*Idealist*) are listed in Table 3. The term ''low'' was removed because it is part of many author names.

The calculation of the sensitivity-precision products led to a new order of terms. The consecutive connection of these terms with the Boolean operator ''OR'' produced the final set of search strategies. Their performance is shown in Table 3.

After the addition of the term ''diagnos*,'' the sensitivity in our test set reached 100%. Every additional term then only produced a decrease of retrieval precision. The combination of the first six terms resulted in a sensitivity of 91.8% (95% confidence interval [CI] 81.9 to 97.3%) and an NNR of 10.9 (95% CI 8.5 to 14.3). That is, almost eleven articles have to be read to identify one on diagnostic accuracy. This strategy seemed to be a good compromise with a high sensitivity and a reasonable precision.

By adding the two terms ''negative'' and the truncated expression diagnos* to search strategy 6, sensitivity reaches 100% (95% CI 94.1 to 100%) and an NNR of twenty-seven (95% CI 21.0 to 34.8). This strategy could be appropriate for systematic reviews.

The combination of the two truncated terms sensitiv* or detect* resulted in a sensitivity of 73.8% (95% CI 60.9 to 84.2%) and a NNR of 5.7 (95% CI 4.4 to 7.6). This latter strategy might be useful for busy clinicians, who may be interested in achieving reasonable sensitivity while avoiding sifting through hundreds of papers.

Figure 1 provides the detailed search strategies for

Table 3
Development of eight search strategies with stepwise adding of terms

Ranking	Added terms	Search strategy	Summary performance sensitivity (%)	Summary performance precision (%)	Number needed to read (NNR)
1	sensitiv*	Strategy: sensitiv*	45.9	27.7	3.6
2	detect*	Strategy: 1 or detect*	73.7	17.6	5.7
3	accura*	Strategy: 2 or accura*	85.2	14.2	7.0
4	specific*	Strategy: 3 or specific*	86.9	10.4	9.6
5	reliab*	Strategy: 4 or reliab*	90.2	10.4	9.6
6	positive	Strategy: 5 or positive	91.8	9.2	10.9
7	negative	Strategy: 6 or negative	91.8	8.5	11.8
8	diagnos*	Strategy: 7 or diagnos*	100.0	3.7	27.0

Terms were ranked according to their sensitivity*precision product. The number needed to read figure shows how many articles have to be read to identify one on diagnostic accuracy and is equivalent to 1/precision.

three commonly used EMBASE interfaces for a reasonably precise and a comprehensive search strategy.

DISCUSSION

In contrast to Haynes and coworkers [14], we included diagnostic articles published in the comment, correspondence, and editorial sections to increase the likelihood of estimating precision correctly. Additionally, we focused on the most relevant indices, that is, sensitivity and precision, ignoring the less useful parameters of specificity and accuracy. In contrast to Haynes and coworkers [15], we did not find any advantage of combining EMTREE terms with text words.

For example, a search with the EMTREE term ''DIAGNOSIS # (explode)'' achieved 93.7% sensitivity and 4.0% precision (NNR = 25). The addition of text words with the Boolean operator ''OR'' would increase the sensitivity but at the cost of worse precision. In our search strategy, however, searches with sensitivities of about 90% were associated with precision of about 10%.

Our aim was to build useful search strategies for systematic reviews requiring very high sensitivity. The precision of search strategies, however, is important for busy health professionals but cannot be fully neglected in systematic reviews either. Search strategies should be evaluated in a context of time investments (cost) and consequences (of missing useful papers) [16]. In analogy to the assessment of the impact of language restrictions on summary measures in systematic reviews [17], the impact of using search strategies with lower sensitivity and higher precision on the summary measures in diagnostic reviews could be evaluated. Finally, this method could be applied to build optimal search strategies to detect diagnostic accuracy research for MEDLINE.

Figure 1
Description of search strategy syntax for three commonly used interfaces

EMBASE interface	Search syntax	
	Specific search	Comprehensive search
Datastar	sensitiv$ detect$	sensitiv$ detect$ accura$ specific$ reliab$ positive negative diagnos$
Ovid	sensitiv$ or detect$	sensitiv$ or detect$ or accura$ or specific$ or reliab$ or positive or negative diagnos$ or di.fs.
SilverPlatter	sensitiv* or detect*	sensitiv* or detect* or accura* or specific* reliab* or positive or negative diagnos*

In our study, we hand searched four important general medical journals for the year 1999 to find diagnostic studies. The restriction to four important general medical journals might limit the generalizability of the search strategy. The restriction to 1999 was reflected in the width of the confidence intervals for sensitivity. We did not measure the test/retest reliability of our final strategies. Independent reassessment of the search performances in another set of gold standard articles would be useful.

Some terms such as ''low'' had to be removed from further analysis. Had we been able to analyze frequencies not by words, but by phrase, we would have been able to identify the proportion of those terms that were used in the diagnostic context (e.g., low accuracy). Including those terms could have been potentially relevant for the filter.

CONCLUSION

The identified search terms allow the choice of either reasonably sensitive or reasonably precise search strategies for the detection of diagnostic accuracy studies in EMBASE. This strategy is useful both for busy health care professionals who value precision and for reviewers who value sensitivity. In practice, clinicians combine the strategies proposed here with terms indicating the specific area of interest, very often a particular disease.

ACKNOWLEDGMENT

We thank Julie Glanville, from the Centre for Reviews and Dissemination in York, United Kingdom, for providing the word frequency analysis and for commenting on an earlier draft.

REFERENCES

1. 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 Nov–Dec;1(6):447–58.

2. Glanville J. Identification of research (phase 3): conducting the review (stage II). [Web document]. In: Undertaking systematic reviews of research on effectiveness. CRD report no. 4. 2001;3–11. [rev. Mar 2001; cited 22 Jun 2002]. <http://www.york.ac.uk/inst/crd/report4.htm>.

3. Jadad AR, McQuay HJ. A high-yield strategy to identify randomized controlled trials for systematic reviews. Online J Curr Clin Trials 1993 Feb 27;Doc No 33:3973 words.

4. Boynton J, Glanville J, McDaid D, Lefebvre C. Identifying systematic reviews in MEDLINE: developing an objective approach to search strategy design. Information Science 1998 Summer;24(3):137–54.

5. 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 Spring;15(2):281–96.

6. Deville WL, Bezemer PD, Bouter LM. Publications on diagnostic test evaluation in family medicine journals: an optimal search strategy. J Clin Epidemiol 2000 Jan;53(1):65–9.

7. McKibbon KA, Walker Dilks CJ. Beyond ACP Journal Club: how to harness MEDLINE for diagnostic problems. ACP J Club 1994 Sep–Oct;(121Suppl 2):A10–2.

8. Irwig L, Tosteson AN, Gatsonis C, Lau J, Colditz G, Chalmers TC, Mosteller F. Guidelines for meta-analyses evaluating diagnostic tests. Ann Intern Med 1994 Apr 15; 120(8):667–76.

9. Cochrane Methods Working Group on Systematic Review of Screening and Diagnostic Tests. Screening and diagnostic tests: recommended methods. [Web document]. [rev. 8 Feb 1998; cited 22 Jun 2002]. <http://www.cochrane.org/cochrane/sadtdoc1.htm>.

10. Haynes, op. cit.

11. National Library of Medicine. PubMed. [Web document]. Bethesda, MD: The Library, 2001. [rev. 10 Oct 2001; cited 22 Jun 2002]. <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi>.

12. Embase. [Web document] [rev. 1 Apr 2000; cited 22 Jun 2002] <http://library.dialog.com/bluesheets/html/bl0072.html/>.

13. Boynton, op. cit.

14. Haynes, op. cit

15. Ibid.

16. Jadad, op. cit.

17. 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 Aug 2;350(9074):326–9.

Received July 2002; accepted December 2002