Techniques to perform a bibliography research

Técnicas para la realización de una búsqueda bibliográfica

J. Navarro
Health Department of Valencia-Physician-Malvarrosa. Valencia

Abstract

The growth in biomedical literature forces us to be more and more rigorous with precise data research. In fact, no bibliography database covers total pertinent data. However, it offers millions of references related to our topic of interest, which may apparently look sufficient. Every strategy for research bibliography should be organized around some key steps. Firstly, formulating a research question (usually therapeutically) and transfer it to a normalized well-documented language. Secondly, selecting proper data resources and combine initial research computer with a subsequent research handbook. Thirdly, formulating a research profile based on logical links adopted from Boolean language. Finally, correcting the strategy by reformulating if its precision was not high. To facilitate the elaboration of a precise strategy there are limit fields, thematic filters and methodological or quality filters which allow specifying data on diagnostic proceeds, therapeutic interventions, etiology or prognosis. In addition, the Cochrane Collaboration offers strategies to make systematic revisions.

Keywords: bibliography, research, strategy.

Introduction

The scientific-medical literature has been threatened by three phenomena: the obsolescence, the dispersion and the excess of information. The growth of the periodical publications is not able to absorb the own growth of the investigating activity and the subsequent offer of manuscripts. Likewise, the secondary literature (indexes, summaries, bibliography) is forced to grow at a rhythm equal to the one experimented by the primary literature. Any bibliography search about a specific subject shall show a high concentration of references in a reduced number of journals, together with a remarkable dispersion of the rest of references in numerous journals. This search process is still more complex if we remember that together with the diversity of documental supports (prints, CDs, online) there is the so-called “grey literature” or opaque, of difficult commercial access (investigation reports, pre-publications, congress resumes and doctoral thesis).

In order to avoid the access to the already outdated information or of scarce relevance, data recovery systems (DRS) have been developed. At present, the access is done through databases, managed by distributors, some of them free. There are also solicitors of bibliography databases (as Procite or Reference Manager) that make possible the complete recovery of information, as well as later information treatment, included the bibliometric analysis.

The main biomedical indexes are the Index Medicus (published by the National Library of Medicine of the United States), Excerpta Medica (published by Excerpta Medica Foundation in Amsterdam) and Current Contents and Science Citation Index (both published by the Institute of Scientific Information of Philadelphia). Their most known databases are MEDLINE (Index Medicus), EMBASE (Excerpta Medica) and SciSearch (Science Citation Index and part of Current Contents). MEDLINE (popularized by PubMed) and EMBASE are the most used databases, as they...
contain almost 100% of the biomedical publications, complementing with: EMBASE covers the 72% of the total of biomedical journals and MEDLINE the 65%, overlapping in a 35%.

Structure of a bibliographic reference
The bibliographic search is supported in the correct delimitation of the different parts of a bibliographic reference. Each record is divided into categories of information named field. The fields LA (language), PT (publication type) and PY (publication year) are known as “limit fields” as they allow a more precise recovery. Most of the fields are of fixed length. The basic fields are: abstract (ab), address (ad), author (au), country of publication (cp), language of article (la), publication type (pt), publication year (py), medical headings (mh), source (so) and title (ti). Table 1 depicts a reference example and table 2 depicts the description of MEDLINE of the quoted reference according to the fields.

Phases of the bibliographic search
Any non-structured search is a frivolity that might entail a great loss of time. Only MEDLINE contains more than 100 million references. If too many records appear, one can talk about “noise” and if there not many, of “silence”. Our objective is the precision. For this, it is advisable to follow the steps that we describe below.

Subject delimitation
An adequate protocol of bibliographic search starts with a clinical question. Such clinical question can be of six types (etiology, prevention, diagnosis, prognosis, treatment or economic impact), each of them requires different sources and strategies. The investigation question should be translated into a documental language, preparing a profile or search equation: i.e. identifying the specific terms and combining them adequately in a formula that the DRS of the used database can decipher. For example, MEDLINE has a thesaurus named MeSH (medical subject headings) that differentiates among describers (words that express the conceptual content of the document), principal (represent the principal matter, they appear below the acronym MJME (major MeSH) and are marked with an asterisk) and secondary (matters related indirectly, they appear under the acronyms MIME ([minor MeSH] and are marked with an asterisk). Subject delimitation

Selection of sources
Due to its great popularity, it has to be stressed that an exclusive access to MEDLINE is not sufficient (with a perfectly designed strategy), we would find the 65% of the pertinent information. Turning to MEDLINE, EMBASE and Current Contents, and carrying out afterwards a manual search (in the bibliography of the located articles), possibly we are able to find almost 100% of the pertinent references.

Formulation of the search profile
According to the Boolean language, logical intersection links are known as union and exclusion of the expressions AND, OR and NOT respectively. The first one looks for what is common of two sub-sets; the second looks for everything what is understood in two sub-sets and the third one includes only one sub-set, deleting another sub-set. There are also syntactic links (WITH, NEAR, etc.) and the so-called search wild cards (“?” and “*”). Figure 1 depicts a diagram with “diabetes” disease in order to appreciate more the Boolean operators according to the diagrams of Venn.

Strategy correction
The precision index (number of pertinent articles from the total of located articles) allows us knowing if the search profile was adequate. It has to be higher than 80%. If negative, the strategy has to be ruled out and it has to be reformulated. Table 3 depicts the formula to estimate the precision index.

Table 1. Example of bibliographic reference

Table 2. Description of MEDLINE about the reference of table 1 according to the fields
- **TI**: Effect of intensive control glucose on cardiovascular outcomes and death in patients with diabetes mellitus: a meta-analysis of randomized controlled trials.
- **AB**: [The abstract comprehends the sections: background, methods, findings, interpretation and funding. It is not included due to its extension]
- **AD**: Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK, kkr25@medschl.cam.ac.uk
- **AU**: Ray KK
- **AU**: Seshasai SR
- **AU**: Wijesuriya S
- **AU**: Sivakumaran R
- **AU**: Preiss D
- **AU**: Erqou S
- **AU**: Sattar N
- **LA**: eng
- **GR**: British Heart Foundation/United Kingdom
- **PT**: Journal Article
- **PT**: Meta-Analysis
- **PT**: Research Support, Non-US Gov’t
- **PL**: England
- **TA**: Lancet
- **MH**: Blood Glucose/*metabolism
- **MH**: Cardiovascular Disease/epidemiology/etiology/*prevention & control
- **MH**: *Diabetes Mellitus, Type 2/blood/complications/mortality/*prevention & control
- **SO**: Lancet. 2009;373(9677):1765-72

Table 3. Example of the formula to estimate the precision index
\[ \text{Precision index} = \frac{\text{Number of pertinent articles}}{\text{Number of located articles}} \]
Figure 1. Example about diabetes. We want to carry out a search that correlates diabetes and hypertension. The sub-set A will be “diabetes” and the sub-set B “hypertension”. According to the diagram 1 (reunion), we would find all the references about diabetes or hypertension (diabetes OR hypertension), according to diagram 2 (intersection), we would find the references about diabetes besides hypertension (diabetes AND hypertension); according to the diagram 3 (negation), we would find the references about diabetes but not about hypertension (diabetes NOT hypertension).

Strategy publication

In the edition of any investigation article in which the bibliography is relevant (especially systematic revisions), the used search strategy has to be specified in detail, as well as the update date of each database (in MEDLINE, the UD code [update code] defines the month and the year of update of each reference).

Structure of a search strategy

If the database contains a thesaurus (as MEDLINE), the selection of terms from it is adequate and not through a free search. For example, if we want to locate references about the treatment of diabetic foot ulcer, through MeSH of MEDLINE we obtain the term “Diabetic Foot” as recognized describer in the thesaurus. Likewise, it indicates us the evolution of its indexation and the semantic relations with other terms. In order to have access to the field of the ulcer treatment, we would specify in MEDLINE (Ovid) to find only everything related to the treatment: “Diabetic Foot/th” (in Pubmed, between square brackets: “diabetic Foot[th]”). As we will see later on, “th” is a thematic filter of MEDLINE that corresponds to therapy. “Diabetic Foot” appears inside “Diseases Categories”, in “Nutritional and Metabolic Diseases, Metabolic Diseases, Diabetes Mellitus, Diabetic Neuropathies” and also in “Endocrine Diseases. Diabetes Mellitus. Diabetic Neuropathies”. The term “Diabetic Foot” was indexed between 1966 and 1993 as “Foot Disease” and between 1992 and 1993 as “Foot Ulcer”, having been indexed since 1994 as “Diabetic Foot”.

Once the describers that more specifically depict what we are looking for are chosen, in a way that each of them constitutes a sub-set of references, the logical and syntactic links are chosen adequately. There are links that deputate the search (and, with, near, not, the brackets) and links that broaden it (or, ?, *). It is not the same to put “smoking OR obesity AND diabetes” than “smoking OR (obesity AND diabetes)”. “Smoking”, “obesity” and “diabetes” are three sub-sets of references.

In the first case, the database can randomly perform the first link of “smoking” and “obesity” or the intersection of “obesity” and “diabetes”. In the second case, it performs the intersection of “obesity” and “diabetes” first (due to the fact that it is between brackets”) and finally the new sub-set is added to “smoking”.

With the IN operator we can undertake a search in specific fields. Three are the fields that have a higher specificity in the search: the title (TI), the abstract (AB) and the describers (in MEDLINE: MeSH; in Ovid: SH: if we specify MJME, we recover only the main one (with asterisk), and if we specify MIME, the secondary ones (without asterisk). For example, “diabetes in ti.ab.mesh”.

There are the so-called limit fields: language (LA or LG [Ovid], language), publication year (PY, publication year) and publication type (PT, publication type). This last field is more important as from the evidence-based medicine point of view (EBM) as it allows finding clinical trials, clinical practice guidelines, reviews or meta-analysis, among other types of documents. Many indexes allow elaborating advanced strategies automatically, including the use of limit fields.

Let us consider that we want to know what is published about the diabetic foot in Spanish between 2005 and 2009. We identify the number of searches with the symbol #:

#1 Diabetic Foot/ti,ab,sh
#2 Spanish/lg
#3 2005-2009/py
#4 #1 AND #2 AND #3

Now, in such sub-set, we want to know if there are clinical trials, reviews or meta-analysis:

#5 Clinical trial/pt
#6 Review/pt
#7 Meta-analysis/pt
#8 #5 OR #6 OR #7
#9 #4 AND #8

After finding a sub-set of references about the determined term, MEDLINE counts with thematic filters (subheadings) that allow a higher precision in the search. Table 4 offers several examples.

The EBM, taking in account that it sets out an analysis of the biomedical literature through the evaluation of specific studies or through systematic reviews, suggests the use of “methodological filters” or “quality filters” as adapted search strategies (diagnostic procedures, therapeutic interventions, etiology, prognosis, systematic reviews, controlled clinical trials, clinical practice guidelines).

Table 3. Formula for the estimation of the precision index

<table>
<thead>
<tr>
<th>Precision index =</th>
<th>No. of relevant documents</th>
<th>No. of total documents</th>
<th>× 100</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Advice to specify the search strategies

Diagnostic procedures

The best simple term is “sensitivity in ti,ab,mesh” (in Ovid: “sensitivity. ti,ab,mesh”).

A maximum of sensitivity is obtained with:
- “sensitivity and specificity”
- “sensitivity in ti,ab,mesh”
- “diagnosis in mesh”
- “radionuclide imaging in mesh”
- “diagnostic use in mesh”
- “specificity in ti,ab,mesh”

A maximum of specificity is obtained with:
- “exp sensitivity and specificity”
- “(predictive and value*) in ti,ab,mesh”

Therapeutic interventions

The best simple term is “clinical-trial in pt” (in Ovid: “clinical-trial. pt”).

A maximum of sensitivity is obtained with:
- “randomized-controlled-trial in pt”
- “drug therapy in mesh”
- “therapeutic use in mesh”
- “random* in ti,ab,mesh”

A maximum of specificity is obtained with:
- “(double and blind*) in ti,ab,mesh”
- “placebo” in ti,ab,mesh”

Etiology

The best simple term is “risk in ti,ab,mesh”.

A maximum of sensitivity is obtained with:
- “exp cohort-studies”
- “exp risk”
- “(odds and ratio*) in ti,ab,mesh”
- “(relative and risk) in ti,ab,mesh”
- “(case and control*) in ti,ab,mesh”

A maximum of specificity is obtained with:
- “prognosis in mesh”
- “survival-analysis”

Analysis of simple and complex strategies

Imagine that we want to find articles in English published in 2009 about dyslypidemia and diabetes (in Pubmed the limitation to changes goes between square brackets: “diabetes in ab” would be “diabetes” [ab]):

Search strategy 1
#1 Diabetes in ti,ab,mesh
#2 Hyperlipidaemia in ti,ab,mesh
#3 #1 OR #2
#4 2009 in py
#5 English in la
#6 #3 AND #4 AND #5

First, the terms that might recover the pertinent search subject are chosen (in medical language and many synonyms (diabetes, hyperlipidaemia). MEDILINE facilitates all the possible semantic relations through its thesaurus suggesting us the term that will recover more references. Through the two first searches (#1 and #2) the references are recovered which offer such terms in the fields corresponding to title, abstract and describers (TI, AB and MESH), as they are more specific. Then, in #3 both sub-sets are added through the logical link OR. In #4, the articles comprised between 2000 and 2009 are recovered through the PY field (publication year). In #5, the articles originally published in English are recovered through the LA field (language). Through this field, MEDILINE makes a difference among all the articles in two large groups: “English” and “non English”, besides specifying the original language. Finally, the
intersection is made through the logical link AND of the three obtained sub-sets in #5.

**Search strategy 2**

Te following example tries to recover the reviews about the hypolipemiant treatment in the prevention of chronic complications in diabetic patients:

#1 Diabetes Mellitus. Type 2/blood/complications/mortality/prevention & control
#2 Anticholesteremic Agents
#3 meta-analysis in pt

#4 #1 AND #2 AND #3

**Search strategy 3**

The last example is the methodological filter proposed by the Cochrane Collaboration in order to identify randomized clinical trials (RCT) and its adaptation to the prevention and cardiovascular treatment in diabetic persons.17-21

#1 randomized controlled trial in pt
#2 controlled clinical trial in pt
#3 “randomized-controlled-trials”/all subheadings
#4 random allocation in pt
#5 double-blind method
#6 single-blind method
#7 #1 OR #2 OR #3 OR #4 OR #5 OR #6
#8 clinical trial in pt
#9 explode “clinical-trials”/all subheadings

#10 (clin* near trial*) in ti,ab
#11 (singl* or doubl* or trebl* or tripl*) near (blind* or mask*)
#12 “placebos”/all subheadings
#13 placebo* in ti,ab
#14 random* in ti,ab
#15 “research-design”/all subheadings
#16 #8 OR #9 OR #10 OR #11 OR #12 OR #13 OR #14 OR #15

#17 tg=comparative study
#18 explode “evaluation-studies”/all headings
#19 “follow-up-studies”/all subheadings
#20 “prospective-studies”/all subheadings
#21 (control* or prospectiv* or volunteer*) in ti,ab
#22 #17 OR #18 OR #19 OR #20 OR #21
#23 #7 OR #16 OR #22
#24 (tg=animal) NOT (tg=human)
#25 #23 NOT #24

**Strategy to identify intervention studies (antihypertensive, hypolipemiant and antiaggregant)**

**Antihypertensive**

#31 explode Antihypertensive Agents therapeutic use in mjme
#32 explode Hypertension drug therapy in mjme

#33 #31 OR #32

**Hypolipemiant**

#31 explode Anticholesteremic Agents therapeutic use
#32 explode Hypercholesterolemia drug therapy

#33 #31 OR #32

**Antiaggregant**

#31 explode Aspirin therapeutic use in mjme
#32 explode Platelet Aggregation Inhibitors therapeutic use in mjme

#33 #31 OR #32

**The diabetes problem**

#26 explode Cardiovascular Diseases/prevention and control in mjme
#27 explode Cardiovascular Diseases/mortality in mjme
#28 explode Cerebrovascular Disorders/prevention and control in mjme
#29 explode Cerebrovascular Disorders/mortality in mjme
#30 explode Diabetes mellitus, Insulin Dependent/complications in mjme

#31 #26 OR #27 OR #28 OR #29 OR #30

#32 explode Hypoglycemic Agents therapeutic use
#33 explode Diabetes Mellitus drug therapy

#34 #32 OR #33

#35 #31 AND #34

**Final filter and synthesis**

#25 AND #30 AND #33 AND #34

**Conclusions**

Every biomedical bibliographic information search needs an adequate strategy due to the high number of published articles each year and the risk of not having access to those really interesting ones or “stumble” with too much not pertinent information.

The first step is to formulate a clinical question (normally, therapeutic) and adapt it to the documental language: the normalized terms and adequately combined through logical links.

The selection of bibliographic sources is also relevant. Taking into account the popularity of MEDLINE through the PubMed, it is convenient to stress that the exclusive access to this database has a selection turn that might suggest a loss of precision in our search, though it is of great help.

An advanced search strategy should consider the use of thematic filters of limit fields, being the identification of the types of documents that better respond to the set out questions of great help.
In any search, one has to know how to rectify the time, according to our objective: find quickly relevant information as for example about treatment (efficacy, adverse effects) or to find everything about a matter because we need a complete review (systematic or not).

Declaration of potential conflict of interests

J. Navarro states that there are no conflicts of interest as regards to the content of this article.

References