A major pitfall in the search strategy on PubMed

Ghazi O. Tadmouri, PhD, MSc, Nisrine Bissar-Tadmouri, PhD, MSc.

Bibliographic indexes in biomedicine. Scientists working in the fields of biomedical sciences are exceptionally well served by many high quality databases. Gaining access to these is usually a top priority for any success.[1] Of these databases, PubMed, the on-line version of MEDLINE developed by the United States of America National Center for Biotechnology Information (NCBI) at the National Library of Medicine (NLM), is considered the most significant barrier-free biomedical resource available on the World Wide Web.[2] PubMed provides a strong health discipline indexing coverage and currently catalogues over 12 million biomedical citations in 4,600 journals published in more than 70 countries.[1,3]

The uses of PubMed as a bibliometric tool. PubMed depends on a text-based search that uses an indexing system for rapid retrieval of information. In this system, the citation information is broken into index fields such as journal name, author name, title, primary author's address, language of publication, and others. The power of the PubMed search could then be further enhanced by the use of search rules, syntax, and qualifying terms with search field abbreviations.[1] Hence, PubMed goes beyond the definition of a simple search engine for biomedical citations, since it can also be used with success to perform bibliometric studies to assess different aspects of research outputs[4] such as: 1. Gross research (or publication) output of medical or related subjects among world countries, cities, institutes, and authors. 2. The history of growth and development of biomedical research in specific world locations. 3. Analysis of publication activity and quality by merging journal publication patterns derived from PubMed with markers of publication citation, such as the Science Citation Index Impact Factor and Immediacy Index. 4. Studying the objectives of biomedical research according to time or geography by detailed examination of the themes and keywords encountered in PubMed.[4]

Geography of biomedical publications in Arab countries. In the last few years, the geography of biomedical publications became the subject of detailed analyses.[5,6] Many of these studies relied solely on PubMed for data collection.[7-14] In Arab countries, one study focused primarily on the Gulf Corporation Council member countries, and demonstrated that the Kingdom of Saudi Arabia (KSA) occupies a leading position in terms of biomedical publications.[15] Two subsequent analyses confirmed this conclusion, but revealed drastic contradictions in the data related to some other Arab countries.[4,16] That is why we aimed at a detailed investigation of this issue with an emphasis on the effect of the languages used in reporting the affiliations of Arab authors on the statistical results drawn from bibliographic databases.

Non-sensitive versus sensitive search strategies on PubMed. As in our earlier study,[4] non-sensitive and sensitive search strategies, including the names of Arab countries as well as their variants in several world languages, were directed to PubMed within a single hour limit (May 20th, 2003). Non-sensitive queries used in this study included simple syntaxes, such as: Qatar[affiliation], in which the tag [affiliation] has the function of collecting all published articles carrying the country name 'Qatar' in the affiliation field.

Our 5-year experience on the evolution of biomedical research in Arab countries[4,17] allowed us to compile a dictionary of highly sensitive search strategies for citations from Arab countries. The implementation of these sensitive strategies offers the advantage of collecting citations missed by non-sensitive search strategies and automatically avoiding false-positive citations like those from the affiliation "Jordan", encountered in the United Kingdom or Northern Ireland, and the affiliations "Lebanon" and "Palestine" that are city names in the United States. With the exception of Qatar and

From the Department of Biology, Bioinformatics Unit, Fatih University, Istanbul, *Turkey*.

Address correspondence and reprint request to: Prof. Ghazi O. Tadmouri, Department of Biology, Bioinformatics Unit, Fatih University, 34500 Buyukcekmece, Istanbul, *Turkey*. Tel. +90 (535) 8232219. Fax. +90 (212) 8890832. E-mail: tadmouri@hotmail.com

Djibouti, retrieving data from PubMed for the rest of the Arab countries requires careful implementation of sensitive queries (**Table 1**).

An example of a sensitive strategy to search for articles authored by affiliates of Palestinian institutes would have the following syntax: Palestine[affiliation] NOT TX[affiliation] NOT Jordan[affiliation] NOT Lebanon[affiliation] NOT Egypt[affiliation]. In this query, the 7 'NOT' Boolean operators would exclude false-positive Palestinian data. In extreme cases, highly sensitive queries require the use of at least 19 exclusions for Lebanon and 27 exclusions for Jordan (**Table 1**).

In the framework of this study, data of Arab citations for years 1987 through to 2001 were extracted from PubMed, transferred to a local database analysis system, and then inspected for any inconsistencies. The obtained citation results were then subdivided into 3 main groups depending on whether they are written in English, French, or in other languages.

When non-sensitive and sensitive search strategies were translated into PubMed format, the difference in the total number of citations received was noted. Compared to the non-sensitive strategies, sensitive strategies retrieved accurate results regarding corresponding author affiliations, thereby minimizing the potential for bias caused by false-positives (**Table 2**).

Language analysis for biomedical citations from Arab countries. Our results demonstrate that Arab biomedical scientists do not publish their findings exclusively in English journals. An important percentage of publications do also appear in French journals. The ratio of French to English citations is 1:4 in Lebanon, 1:1 in Algeria, and goes up to 2:1 in Tunisia, Morocco, and Mauritania (**Table 2**). To a lesser extent, articles authored by biomedical scientists affiliated to Arab institutes also appeared in different languages such as Italian, Spanish, German, Arabic, Russian, Slovakian, Czech, Portuguese, Swedish, Norwegian, Bulgarian, and Serbian/Croatian (**Table 2**). Another important observation is that even in English journals, authors from the Arab Maghreb countries and Lebanon prefer to report their addresses in French instead of English (data not shown). This is in support of a study, which demonstrated that 91% of PubMed-indexed medical articles reported by primary Tunisian authors from 1965 to 1999 have been published in French.[18]

Pitfalls in search strategies on PubMed. On-line literature searches of bibliographic databases such as PubMed are now integral to the lives of many scientists. The recently added features to the latest release of PubMed allow sensitive gleaning from a basic search, while advanced functions offer great flexibility, speed, and focus.[1] However, these advantages may be easily degraded if the user does not utilize proper keywords or implement correct syntaxes. In this case, the obtained results may be highly biased and sometimes misleading.[19]

The history of Arabs and the cultural ties between south and north Mediterranean countries continue to cast a major influence on the language of communication among many Arab scientists, whether English, French, or others. For this, reason, some Arab scientists prefer to report their addresses in French irrespective of the language of the hosting journal.[18] This is in addition to the inconsistencies in reporting the correct addresses of the institutes to which they are affiliated.[4]

In a recent study, data on the current situation of biomedical science were under represented for countries such as Tunisia, Algeria, and to a lesser extent for Lebanon.[7] These countries are well known for the advanced, and dynamic scientific research institutes they have, such as Institut Pasteur in Tunisia and the American University of Beirut Medical Center in Lebanon. In the aspect of language of communication, the common factor that may group these countries is the influence of the French language on education and science. Such an influence, if not taken into consideration, may have a great impact on the resulting data when the search for Arab PubMed-indexed citations is conducted. This could explain the high similarity between the data extracted by non-sensitive search strategies in our present study (**Table 2**) and the overall data reported by Shaban and Abu-Zidan.[16] However, by implementing sensitive search strategies, considering false-positive addresses and all possible languages used for reporting author affiliations, countries of the Arab Maghreb and Lebanon easily take their places among the most biomedically proliferating Arab countries following KSA, Egypt, and Kuwait (**Table 2**).

A study investigating specialist biomedical journals demonstrated that errors in key elements of references, such as 'address' or 'affiliation', are prevalent.[20] When such 'erroneous' or 'inconsistent' address information is faithfully passed from the author to the journal publisher, and ultimately to PubMed,[1] the use of the 'affiliation field' becomes questionable since it becomes more difficult, time consuming, and frustrating for readers and librarians to identify all possibly published papers in a certain country. Such problems were frequently encountered when collecting data for biomedical publications in the United Kingdom,[21] KSA,[4] Lebanon (data not shown), and Turkey.[22]

Recommendations. The use of precise sensitive strategies in the present study limit the amount of non-relevant citations retrieved and lowers the amount of work in reviewing citations for eligibility. It is our hope that these highly sensitive search strategies will be of great help for reviewers working on the statistical evaluation of biomedical publications in Arab countries. However, despite this high sensitivity it should be noted that the possibility of a minimal error

Table 1 - Non-sensitive and sensitive search strategies in PubMed format used to extract Arab citations.

Country	Non-sensitive query	Sensitive query
Algeria	Algeria[affiliation]	Alger*[affiliation] NOT Leeds[affiliation] NOT USA[affiliation] NOT Spain[affiliation]
Bahrain	Bahrain[affiliation]	Bahrain[affiliation] OR Bahrein[affiliation]
Comoros	Comoros[affiliation]	Comor*[affiliation] NOT Canada[affiliation] NOT France[affiliation] NOT Italy[affiliation] NOT USA[affiliation] NOT Michigan[affiliation]
Djibouti	Djibouti[affiliation]	Djibouti[affiliation]
Egypt	Egypt[affiliation]	Egypt[affiliation] OR Egypte[affiliation]
Eritrea	Eritrea[affiliation]	Eritrea[affiliation] OR Erythree[affiliation] NOT Italy[affiliation] NOT Milano[affiliation] NOT Germany[affiliation]
Iraq	Iraq[affiliation]	Iraq[affiliation] OR Irak[affiliation]
Jordan	Jordan[affiliation]	Jordan*[affiliation] NOT USA[affiliation] NOT UK[affiliation] NOT Ireland[affiliation] NOT Dundee[affiliation] NOT Portugal[affiliation] NOT Bregenz[affiliation] NOT Croatia[affiliation] NOT Germany[affiliation] NOT Spain[affiliation] NOT Poland[affiliation] NOT Nablus[affiliation] NOT Netherlands[affiliation] NOT Ulster[affiliation] NOT Canada[affiliation] NOT Edinburgh [affiliation] NOT Australia[affiliation] NOT Stanford[affiliation] NOT Strathclyde[affiliation] NOT Glasgow[affiliation] NOT Manchester[affiliation] NOT Plymouth[affiliation] NOT Zagreb [affiliation] NOT Kenneth[affiliation] NOT Frankfurt[affiliation] NOT Alabama[affiliation] NOT France[affiliation] NOT Israel[affiliation]
Kuwait	Kuwait[affiliation]	Kuwait[affiliation] OR Koweit[affiliation]
Lebanon	Lebanon[affiliation]	Lebanon[affiliation] NOT Bronx[affiliation] NOT USA[affiliation] NOT Palmyra[affiliation] NOT Jersey[affiliation] NOT Hampshire[affiliation] NOT Pennsylvania[affiliation] NOT Annville [affiliation] NOT Wisconsin[affiliation] NOT 03766[affiliation] NOT 17042[affiliation] NOT Oregon[affiliation] NOT NJ[affiliation] NOT Dartmouth[affiliation] NOT Hitchcock[affiliation] NOT Koala[affiliation] NOT Codman[affiliation] NOT 03756[affiliation] NOT "Magnetic Imaging" [Affiliation] NOT Spain[affiliation] OR Liban[affiliation]
Libya	Libya[affiliation]	Liby*[affiliation] NOT China[affiliation]
Mauritania	Mauritania[affiliation]	Mauritani*[affiliation]
Morocco	Morocco[affiliation]	Morocco[affiliation] OR Maroc[affiliation]
Oman	Oman[affiliation]	Oman[affiliation] NOT USA[affiliation] NOT Sweden[affiliation] NOT Stockholm[affiliation]
Palestine	Palestine[affiliation]	Palestine[affiliation] NOT Jordan[affiliation] NOT TX[affiliation] NOT Lebanon[affiliation] NOT Egypt[affiliation]
Qatar	Qatar[affiliation]	Qatar[affiliation]
Saudi Arabia	"Saudi Arabia"[affiliation]	"Saudi Arabia"[affiliation] OR "Arabie Saoudite"[affiliation] OR KSA[affiliation] NOT Oman [affiliation]
Somalia	Somalia[affiliation]	Somali*[affiliation] NOT Djibouti[affiliation]
Sudan	Sudan[affiliation]	Sudan[affiliation] OR Soudan[affiliation] NOT Kenya[affiliation] NOT Nigeria[affiliation] NOT France[affiliation]
Syria	Syria[affiliation]	Syri*[affiliation] NOT Finland[affiliation] NOT India[affiliation] NOT England[affiliation]
Tunisia	Tunisia[affiliation]	Tunisi*[affiliation]
United Arab Emirates	"United Arab Emirates"[affiliation]	"United Arab Emirates"[affiliation] OR UAE[affiliation] OR "Emirats Arabes Unis"[affiliation]
Yemen	Yemen[affiliation]	Yemen[affiliation] NOT USA[affiliation]

Table 2 - Non-sensitive and sensitive search results of PubMed-indexed citations for Arab countries based on the search strategies listed in Table 1 (1987-2001).

Country	Non-sensitive query		Sensitive query					
	All languages		French citations	Other citations	All languages			Percent relative error %
	Citations	Average/ year			Citations	(%)	Average/ year	
Saudi Arabia	6670	445	3	15	6719	(27.5)	448	+0.7
Egypt	5845	390	18	6	5861	(24.0)	391	+0.3
Kuwait	1958	131	0	6	1958	(8.0)	131	-
Tunisia	460	31	1152	0	1709	(7.0)	114	+73.1
Morocco	508	34	1101	0	1695	(6.9)	113	+70
Lebanon	2903	194	254	1	1387	(5.7)	92	-109.3
Jordan	1296	86	0	3	1146	(4.7)	76	-13.1
United Arab Emirates	873	58	0	1	980	(4.0)	65	+10.9
Sudan	604	40	1	1	612	(2.5)	41	+1.3
Oman	501	33	0	2	496	(2.0)	33	-1
Algeria	160	11	203	1	418	(1.7)	28	+61.7
Iraq	377	25	2	2	377	(1.5)	25	-
Libya	287	19	0	5	309	(1.3)	21	+7.1
Qatar	208	14	0	0	208	(0.8)	14	-
Bahrain	205	14	0	0	205	(0.8)	14	-
Syria	106	7	8	8	127	(0.5)	8	+16.5
Yemen	81	5	3	1	80	(0.3)	5	-1.2
Mauritania	11	1	30	0	46	(0.2)	3	+76.1
Palestine	52	3	0	0	45	(0.2)	3	-15.6
Somalia	24	2	1	1	45	(0.2)	3	+46.7
Djibouti	27	2	21	0	27	(0.1)	2	-
Comoros	3	0.2	7	0	11	(0.04)	1	+72.7
Eritrea	110	7	0	0	9	(0.04)	1	-1122.2

could not be ruled out. This is as new false-positive affiliation names may come up in the future. Hence, to continue to be effective and efficient, the sensitive strategies reported herein should be examined periodically to take into account new false-positive addresses indexed on PubMed.

References

1. The National Center for Biotechnology Information (USA). The NCBI Handbook. USA: NCBI; 2003.
2. McEntyre J, Lipman D. PubMed: bridging the information gap. ***CMAJ*** 2001; 164: 1317-1319.
3. Sittig DF. Identifying a core set of medical informatics serials: An analysis using the MEDLINE database. ***Bull Med Libr Assoc*** 1996; 84: 200-204.
4. Tadmouri GO, Tadmouri NB. Biomedical research in the Kingdom of Saudi Arabia (1982-2000). ***Saudi Med J*** 2002; 23: 20-24.
5. Salem S. Bibliometric aspects of medical information in Arab countries. ***Bull Med Libr Assoc*** 1990; 78: 339-344.
6. Hefler L, Tempfer C, Kainz C. Geography of biomedical publications in the European Union, 1990-98. ***Lancet*** 1999; 353: 1856.
7. Moed HF, Van Ark GA, van den Berghe H. Bibliometric indicators of the quality of medical scientific research in The Netherlands and Flanders. ***Ned Tijdschr Geneeskd*** 1995; 139: 1483-1489.
8. Gotzsche PC, Krog JW, Moustgaard R. Bibliometric analysis of Danish medical research 1986-1992. ***Ugeskr Laeger*** 1995; 157: 5075-5081.
9. Bunout D, Reyes H. Biomedical articles by Chilean authors published in international journals in 1997. A review from MEDLINE. ***Rev Med Chil*** 1998; 126: 1132.
10. Favaloro EJ. Medical research in New South Wales 1993-1996 assessed by Medline publication capture. ***Med J Aust*** 1998; 169: 617-622.
11. Rahman M, Laz TH, Fukui T. Health related research in Bangladesh: MEDLINE based analysis. ***J Epidemiol*** 1999; 9: 235-239.
12. Rosselli D. Latin American biomedical publications: the case of Colombia in Medline. ***Med Educ*** 1998; 32: 274-277.
13. Thompson DF. Geography of U.S. biomedical publications, 1990 to 1997. ***N Engl J Med*** 1999; 340: 817-818.
14. Han MC, Lee CS. Scientific publication productivity of Korean medical colleges: an analysis of 1988-1999 MEDLINE papers. ***J Korean Med Sci*** 2000; 15: 3-12.
15. Deleu D, Northway MG, Hanssens Y. Geographical distribution of biomedical publications from the Gulf Corporation Council countries. ***Saudi Med J*** 2001; 22: 10-12.
16. Shaban SF, Abu-Zidan FM. A quantitative analysis of medical publications from Arab countries. ***Saudi Med J*** 2003; 24: 294-296.
17. Tadmouri GO, Bissar-Tadmouri N. Genetic disorders among Arabs as for OMIM™. ***Saudi Med J*** 1999; 20: 4-18.
18. Ben Abdelaziz A, Harrabi I, Aouf S, Gaha R, Ghannem H. Typology of Tunisian medical research indexed in Medline from 1965 to 1999. ***Tunis Med*** 2002; 80: 548-555.
19. 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: 150-153.
20. Sieber R, Holt S. Accuracy of references in five leading medical journals. ***Lancet*** 2000; 356: 1445.
21. Takei N, Verdoux H. Research activities on schizophrenia in 17 non-English speaking countries: A MEDLINE survey. ***Eur Psychiatr*** 1997; 12: 319-320.
22. Gulen R. NeHIR, DoTBar and Deniz: Three biological databases established at Fatih University in Istanbul, Turkey [dissertation]. Istanbul (TR): Fatih University; 2003.