Assessment of bibliographic databases performance in information retrieval for occupational and environmental toxicology

Jean-Francois Gehanno, Christophe Paris, Benoit Thirion, Jean-Francois Caillard

Abstract

***Objective*—To determine the efficiency of the major bibliographic databases by assessing the percentage of references among the total literature available that can be retrieved from each database. We also evaluated the best database combinations to carry out an exhaustive search.**

***Methods*—BIOSIS, EMBASE, MEDLINE, NIOSH-TIC, and TOXLINE were searched on two topics: allergy to latex and asbestos and mesothelioma, in the title, abstract, or keywords (textwords). This search was performed for the years 1994 and 1995. All the records were classified by journal and author's name and were verified for each record whether or not it was indexed in each database. Statistical analysis was performed with χ^2 test.**

***Results*—777 articles in 510 issues were found. The efficiency of each database (percentage of articles recovered) and of combinations varied between 11% and 63% for one database and between 42% and 86% for a combination of two databases. The reasons why these differences exist between databases, and within a database, between two different subjects or two different years are reported.**

***Conclusion*—Firstly, it is not advisable to assert that a bibliography is complete when only one database is searched. Secondly, the efficiency of the databases may be quite different. Finally, it is suggested that the best way to be as exhaustive as possible is to search two or more databases—for example, in EMBASE and TOXLINE, or to a lesser extent EMBASE and MEDLINE. This seems to be the best compromise solution between time consumed for searching and efficiency.**

(*Occup Environ Med* 1998;**55**:562–566)

Keywords: bibliographic databases; efficiency; toxicology

Institute of Occupational Health, Rouen University Hospital, France
J-F Gehanno
C Paris
J-F Caillard

Medical Library, Rouen University Hospital, France
B Thirion

Correspondence to:
Dr JF Gehanno, Institut de Médecine du Travail, Hôpital Charles Nicolle, 1 rue de Germont, F-76031 Rouen Cedex, France. Tel 0033 2 32 888 285; Fax 0033 2 32 888 184; email Jean-Francois.Caillard@chu-rouen.fr

Accepted
25 February 1998

The field of toxicology is vast as it covers numerous different topics—such as drugs, industrial or domestic products, environmental pollutants, and animal or plant toxins. Specialists in occupational and environmental health can be confronted with all of these types of agents and therefore they must be able to retrieve specific information. This information can be found in books, grey literature, and periodicals which are indexed in bibliographic databases. The tremendous growth of medical literature has led to the extensive development of these databases, with CD-ROM and on line automated retrieval systems available since the mid-1980s.[1] The database searches can be made by using keywords to find the information needed, in a minimum amount of time and with a minimum of irrelevant references. However, each database does not index all the existing literature and there are great differences in the number of journals and references indexed and in the indexation system itself.[2] Thus, to evaluate the efficiency of the major bibliographic databases in the field of occupational and environmental toxicology, all the databases were searched with two toxicology topics and the results of the databases alone and in various combinations were compared. Furthermore, we determined the optimal combination of databases to achieve the best compromise between an exhaustive search and searcher's time.

Methods

In October and November 1996 we searched databases on two different subjects: mesothelioma and abestos (search A), and allergy to latex (search B). For MEDLINE and EMBASE, we used keywords from their thesauruses (respectively medical subject headings (MeSH) and Emtree)—that is, asbestos and mesothelioma on the one hand, and hypersensitivity and latex on the other. For BIOSIS, TOXLINE, and NIOSH-TIC which do not have a thesaurus, the keywords used were asbestos and mesothelioma for one search, and hypersensitivity, allergy, and latex for the other. The keywords were searched as textwords (abstract plus title plus keyword fields), and they were combined with the Boolean expressions, according to the recommendations of the National Library of Medicine in USA (NLM, Bethesda, MD, USA), available on NLM's website (URL: http://www4.ncbi.nlm.nih.gov/PubMed/syntax.html/).

Textwords were combined with the Boolean operator "Or" for hypersensititvity Or allergy, which found all the references containing the word hypersensititvity plus all those containing the word allergy, and the Boolean operator "And" for hypersensitivity Or allergy And latex and asbestos And mesothelioma. This permitted the selection of only the references which contained, for example, the word asbestos and the word mesothelioma. The search was made for the years 1994 and 1995. No restriction was

Table 1 Results of the searches for each database or combination

	Total search A		*Total search B*		*Total*	
	n	*%*	*n*	*%*	*n*	*%*
N	66	21	52	11	118	15
M	199	63	173	37	372	48
B	153	49	221	48	374	48
E	188	60	237	51	425	55
T	188	60	267	58	455	59
MN	208	66	193	42	401	52
BN	175	56	245	53	420	54
EN	206	65	247	53	453	58
TN	199	63	283	61	482	62
MT	232	74	321	69	553	71
TB	224	71	334	72	558	72
MB	244	77	325	70	569	73
EM	270	86	304	66	574	74
EB	248	79	358	77	606	78
ET	270	86	384	83	654	84
MTN	241	77	332	72	573	74
TBN	234	74	347	75	581	75
EMN	276	88	312	68	588	76
MBN	253	80	338	73	591	76
EBN	259	82	363	79	622	80
MTB	251	80	378	82	629	81
ETN	276	88	388	84	664	85
EMT	293	93	412	89	705	91
EMB	303	96	414	90	717	92
ETB	291	92	432	94	723	93
MTBN	260	83	389	84	649	84
EMTN	299	95	416	90	715	92
EMBN	309	98	419	91	728	94
ETBN	297	94	436	94	733	94
EMTB	309	98	458	99	767	99
EMTBN	315	100	462	100	777	100

B = BIOSIS; E = EMBASE; M = MEDLINE; T = TOXLINE; N = NIOSHTIC.

made on the language but we eliminated grey literature and books from the initial results of the search to retain only the journals.

The databases used were:

BIOSIS

This is the automated version of *Biological Abstracts*, which provides information on biological sciences and contains more than nine million records from 9000 national and international journals and periodicals (28% published in the United States and 38% in western Europe). It covers the period from 1970 to the present, and increases by about 540 000 records a year. BIOSIS is available, both on CD and on line access, through Ovid (Ovid Technologies, London, UK; URL http://www.ovid.com/db/order/html) or SilverPlatter (SilverPlatter Information, London, UK; URL: http://www.silverplatter.com/offices.html).

EMBASE

This corresponds to the printed *Excerpta Medica* series. This database concerns all the fields of medicine and indexes 3500 journals, among which 55% come from Europe. It contains more than five million references from 1974 to present. EMBASE is available, on both CD and on line access, through Ovid or SilverPlatter.

MEDLINE (MEDLARS ONLINE)

This database is the on line and CD-ROM equivalent of *Index Medicus*, and is produced by the NLM. It contains more than eight million records from over 3500 biomedical national and international journals and periodicals, covering the period from 1966 to the present, and increases by 324 000 records a year. MEDLINE is available, both on CD and on line access, through Ovid or SilverPlatter.

NIOSH-TIC (NATIONAL INSTITUTE FOR OCCUPATIONAL SAFETY AND HEALTH—TECHNICAL INFORMATION CENTER) DATABASE

This database provides information on all aspects of occupational health and safety and is produced by the National Institute for Occupational Safety and Health (Robert A Taft Laboratories, Cincinatti, Ohio, USA). It contains

Table 2 Significance of differences in results between databases or combinations for search A

	B	*E*	*M*	*N*						*Rank*
B										2
E	**									1
M	***	NS								1
N	***	***	***							3
T	***	NS	NS	***						1
	BN	EB	EM	EN	ET	MB	MN	MT	TB	
BN										5
EB	***									2
EM	***	*								1
EN	**	***	***							3
ET	***	**	NS	***						1
MB	***	NS	**	**	*					2
MN	**	**	***	NS	***	***				3
MT	***	NS	***	*	***	*	***			2
TB	***	*	***	NS	***	***	NS	NS		3
TN	*	***	***	NS	***	***	NS	***	***	4
	EBN	EMB	EMN	EMT	ETB	ETN	MBN	MTB	MTN	
EBN										4
EMB	***									1
EMN	NS	***								3
EMT	***	*	*							2
ETB	***	*	NS	NS						2
ETN	*	***	NS	**	**					3
MBN	NS	***	*	***	***	*				4
MTB	NS	***	*	***	***	*	NS			4
MTN	NS	***	***	***	***	***	*	NS		4
TBN	*	***	***	***	***	***	***	**	NS	5

p>0.05; *p<0.05; **p<0.01; ***p<0.001.
χ^2 test, paired cases.
Rank = classification of databases, according to the results and the significance of their differences.

Table 3 Significance of differences in results between databases or combinations for search B

	B	*E*	*M*	*N*						*Rank*
B										2
E	NS									1
M	**	***								3
N	***	***	***							4
T	***	NS	***	***						1
	BN	EB	EM	EN	ET	MB	MN	MT	TB	
BN										5
EB	***									2
EM	***	***								3
EN	NS	***	***							5
ET	***	*	***	***						1
MB	***	**	NS	***	***					3
MN	***	***	***	***	***	***				6
MT	***	*	NS	***	***	NS	***			3
TB	***	NS	NS	***	***	NS	***	NS		2
TN	**	***	NS	*	***	**	***	***	***	4
	EBN	EMB	EMN	EMT	ETB	ETN	MBN	MTB	MTN	
EBN										4
EMB	***									2
EMN	***	***								5
EMT	***	NS	***							2
ETB	***	*	***	*						1
ETN	*	*	***	***	***					3
MBN	*	***	NS	***	***	***				5
MTB	NS	**	***	***	***	NS	***			3
MTN	*	***	NS	***	***	***	NS	***		5
TBN	NS	***	*	***	***	***	NS	***	NS	4

NS p>0.05; *p<0.05; **p<0.01; ***p<0.001.
χ^2 test, paired cases.
Rank = classification of databases, according to the results and the significance of their differences.

about 195 000 records from 160 periodicals and thousands of monographs and technical reports, and covers primarily the period from 1973 to the present. It is available on CD-ROM through the Canadian Center for Occupational Health and Safety (email custserv@ccohs.ca).

TOXLINE

This database reunites the toxicology files originally assembled by the NLM (the CD-ROM equivalent is named TOXLINE Plus) and covers the toxicological literature since 1985. It includes over one million records, with about 170 000 records added each year, with 70 000 contributed by Chemical Abstract Service (CAS), 37 000 by BIOSIS, 12 000 by International Pharmaceutical Abstracts, and 50 000 by the NLM. This database provides references in all areas of toxicology, including occupational toxicology, and is available through SilverPlatter, with CD and on line access.

BIOSIS, EMBASE, MEDLINE, and TOXLINE are also available, only with on line access, through DIMDI (Deutsches Institut für Medizinische Documentation und Information, Köln, Germany).

All the records were classified by journal and author's name. Then, we verified whether or not the records were indexed in each database.

STATISTICAL ANALYSIS

Statistical analysis was performed using χ^2 test for the significance of the differences between databases or combinations (paired cases). Differences were considered significant at $\alpha=0.05$.

Results

Table 1 shows the results. We found 777 different records in 510 issues. This amount of references was estimated to be the whole published literature on the topics studied during 1994 and 1995. The numbers of records provided by each database or combination of databases were compared with this total. Each database provided between 11% (NIOSHTIC) and 63% (MEDLINE) of the total of records for one of the two topics studied. The combination of two databases provided between 42% and 86% of the total records (EMBASE plus TOXLINE on search A) and the combination of three databases provided between 68% and 96% (EMBASE plus MEDLINE plus BIOSIS on search A).

The grading of the "best" databases or combination was achieved with χ^2 test, paired cases (tables 2 and 3).

When just one database was searched, the best results were given by M, E, or T (difference not significant) for search A and by T or E for search B, followed by B in the two cases. When two databases were searched, the best results were given by EM or ET (difference not significant) for search A and by ET for search B, followed respectively by EB, MB, or MT for search A and by EB or TB for search B. When three databases were searched, the best results were given by EMB for search A and by ETB for search B, followed respectively by EMT or ETB (difference not significant) for search A and by EMB or EMT for search B.

Discussion

Grey literature was eliminated from the initial results of our study. However, databases vary widely, not only for periodicals but also for grey literature. For example, BIOSIS indexes some conference proceedings and research reports which represent nearly 35% of its records. This is counterbalanced by the fact that some conference or symposium proceedings are published in special issues or in supplements of

periodicals and are therefore indexed in some databases.

We chose these five databases because they were available worldwide, they contain information on toxicological issues, and they have been adopted by medical libraries.[2] Different versions of these databases are available in a compact disk format and include SilverPlatter, Compact Cambridge, DIALOG OnDisc, and OVID, depending on the database. Furthermore, they cover complementary geographical zones because MEDLINE and NIOSH-TIC are more concerned with North American literature and EMBASE with European literature, whereas BIOSIS and TOXLINE are less well defined, even if their contents primarily originate from English speaking countries.

The amount of references recovered was estimated to be the whole published literature on the topics studied during 1994 and 1995, and the databases were compared with this total. Obviously, this underestimated the real amount of published literature, but we considered that this did not modify our conclusions for several reasons. Firstly, even if numerous local journals, for example in the Asian or Pacific area, are not indexed in these international databases, their availability is limited and articles published in these journals are more often than not of local or regional interest. Furthermore, if an article is innovatory or is an original contribution to knowledge, it will probably be published in an international journal, as international journals require quality articles and the authors attempt to publish in journals indexed in international databases to be quoted in bibliographies of other articles. This is the basis of the Impact Factor system. For example, we performed the same searches, with the French equivalent keywords in a French bibliographic database, INRS-B (provided by the National Institute for Research and Safety), which provided information on all aspects of occupational health and safety. Forty nine records were found (search A plus search B). Among them, 20 were found in this database only and most of them originated from an information periodical for industrial hygienists that contains small and practical articles. The other records came from a French journal, indexed since then in BIOSIS. Secondly, even if numerous local journals are not indexed in these international databases, they all index some international journals from Asia—for example, *Fukuoka Igaku Zasshi* (TOXLINE) or *Chung Hua I Hsueh Tsa Chih Taipei* (MEDLINE and TOXLINE)—or from Eastern Europe—for example *Polski Tygodnik Lekarski* (MEDLINE et TOXLINE), or the *International Journal of Occupational Medicine and Environmental Health* (MEDLINE, NIOSH-TIC). Thus, this contributes to a satisfactory geographical coverage. Thirdly, we considered our search to be exhaustive also because of the retrieval process. Searching in textwords permits a maximum reference retrieval, even if it increases the risk of obtaining irrelevant information. This was not a problem in our study as its purpose was to carry out an exhaustive research.

We chose years 1994 and 1995 to avoid basing our judgment on indexation delay. In some databases, this delay, between the time an article is published and its indexation, can be several months, which depends on the database and the journal concerned.

The comparisons of the results of each search in the databases showed important variations between databases, and, within a database, between two topics, and between two different years. These discrepancies were due to several factors.

The first factor is that databases do not index the same journals, and even if the most important international journals, for example *Lancet*, *Occupational and Environmental Medicine*, or the *American Journal of Industrial Medicine*, are indexed in the five databases studied, numerous journals are indexed in just one database. For example, the journal *Immunology and Allergy Clinics of North America* is not indexed in MEDLINE, whereas EMBASE indexed 12 articles from this journal on the topic of search B, which contributed to the poor results of MEDLINE on this subject compared with EMBASE. This factor partly explains the differences between databases or combinations.

Secondly, some journals are entirely and some are just partly indexed in the databases—that is, some articles in an issue can be indexed whereas some will not be indexed in a database. The journals partly indexed are different between databases. Thus, databases sometimes index the same journals, but not the same articles in an issue of these journals. For example, *Annals of Occupational Hygiene* was indexed in the five databases, but, in Vol 39, number 5, the article by Brown *et al* (p 705–13) was indexed just in EMBASE, the article by Hirst *et al* (p 623–32) was indexed in TOXLINE and BIOSIS, and the article by Rodelsperger and Woitowitz (p 715–25) was indexed in the five databases. This contributes to the differences in results between databases.

Thirdly, indexation of special issues or supplements of periodicals reporting conference or symposium proceedings is different between the databases. For example, number 93 (1 Part 2) of the *Journal of Allergy and Clinical Immunology* reports the 50th Annual Meeting of the American Academy of Allergy and Immunology. Even if this journal is usually indexed by EMBASE, MEDLINE, TOXLINE, BIOSIS, and NIOSH-TIC, this issue only appeared in TOXLINE, BIOSIS, and NIOSH-TIC. Furthermore, according to search B, 40 articles in this issue were indexed in BIOSIS, 23 in TOXLINE, and just one in NIOSH-TIC. This contributes to the differences in results between databases, but also within a database, between two years or two topics. For example MEDLINE provided 63% (199) of the total references of search A, and only 37% (173) on search B. This difference in results was due to the omission of indexation of the issue of the *Journal of Allergy and Clinical Immunology* already mentioned. Another example is that, for NIOSH-TIC, performances

in search A were different between 1994 (25%) and 1995 (16%). This was due to the indexation of the journal *Annals of Occupational Hygiene*, volume 38, number 4, which reports a workshop on the topic (Workshop on Health Risks Associated with Chrysotile Asbestos, Jersey, Channel Islands, 14–17 November 1993). Fourteen articles out of the 26 papers in this issue which were about this workshop were indexed in NIOSHTIC from this issue. On the other hand, there was no special issue indexed in this database in 1995, which contributes to the difference in results.

Fourthly, different databases use different keyword systems—for example the NLM's medical subject headings, and apply them according to different principles.[3] Thus, an article may be indexed in several databases but be retrieved only in one of the databases according to the keywords used in the search and those used by the indexers of the database. This factor was not a problem in our study as we chose only the keywords which gave the highest number of results on each topic. Furthermore, we searched in textword—that is, in titles, abstracts, and keywords used for indexation—which permits the maximun recoverage. Searching by textword can also supplement a search by keywords from the thesaurus, especially when the search gives no articles, or too few.[4]

Nevertheless, this factor must be taken into consideration when a bibliographic search is carried out, particularly when textwords or keywords different from those of the thesaurus are used. This is why it is always advisable to use the keywords provided by the thesauruses of the databases (when they exist). When the thesauruses are hierarchical (in MEDLINE, EMBASE, and TOXLINE), it is often of interest to use the "explode" function as it provides more references because it recovers all the references including the term searched and all the terms which are more specific.[4] For example, in the MeSH, exploding the term asbestos will select the references including the terms asbestos, amphiboles, amosite, crocidolite, and serpentine.

With the reservations already mentioned, we can try drawing up optimal retrieval strategies, based on the two searches achieved. It seems that when just one database is searched, TOXLINE is probably the most interesting for it provides the highest percentage of records among the total available literature in the field of occupational and environmental toxicology, which is in agreement with the structure and origin of this database. Nevertheless, the differences in results between TOXLINE and EMBASE were not found to be significant. BIOSIS and MEDLINE came in second position. More exhaustive results can be performed when combining two databases as we can obtain more than 80% of the total available literature. Moreover, the combination of two databases limited the risks of not recovering a special issue of a journal dealing with the subject of our search, and reduced the effects of the differences in keyword systems between databases, and thus limited the risk of missed information. The ideal combination seems to be TOXLINE plus EMBASE, which also permits a satisfactory coverage of European publications. The results given by this combination were statistically better than those given by the other combinations of two databases. It is difficult to assert whether the best efficiency of these combination in this study was likely to be relevant to other topics in occupational and environmental toxicology because of the small sample of topics and years covered. Nevertheless, in this study, the results between search A and search B were concordant. Furthermore, when searching for adverse drug reactions, Biarez *et al* have also shown that the combination of TOXLINE plus EMBASE is useful.[5]

The combination of TOXLINE, EMBASE, and BIOSIS or EMBASE, MEDLINE, and BIOSIS provide >90% of the available literature. The first combination gave statistically better results in search B and the second in search A.

At present, >80 000 periodicals exist in the world,[6] some dealing with occupational and environmental toxicology.[7] As we are confronted with this considerable amount of information, it is of major importance to have reliable and efficient information retrieval systems at our disposal. Bibliographic databases meet this requirement provided we know their limits. Thus, it is not advisable to assert that a bibliography is complete when only one database is searched. Furthermore, the efficiency of the databases may be quite different and it is important to choose the ones that are best suited to the subject of the study and to combine two or three databases to achieve the best compromise solution between the time taken to search on the one hand and efficiency and quality of the search on the other. When cost is not taken into consideration,[8] the best solution is a multibase search, which can be achieved through DIMDI, DIALOG, or DATA STAR.

We thank Mr R Medeiros and Mrs Y Autain for their advice in editing the manuscript.

1 Cox JJ, Dawson KJ, Hobbs KEF. The electronic information revolution and how to exploit it. *Br J Surg* 1992;**79**:1004–10.
2 Ludl H, Schope LH, Mangelsdorf I. Searching for information on toxicological data of chemical substances in selected bibliographic databases. Selection of essential databases for toxicological researches. *Chemosphere* 1996; **32**:867–80.
3 Hersh WR, Greenes RA. Information retrieval in medicine: state of the art. *MD Comput* 1990;7:302–11.
4 Greenhalgh T. The Medline database. *BMJ* 1997;**315**:180–3.
5 Biarez O, Sarrut B, Doreau CG, *et al*. Comparison and evaluation of nine bibliographic databases concerning adverse drug reactions. *Drug Intelligence and Clinical Pharmacy* 1991;**25**:1062–5.
6 Wolf-Terroine M, ed. Répertoire international des banques de données biomédicales 1991–2. Paris: FLA Consultants, 1991.
7 Takala J. CD-ROMs and databases as vehicles for chemical safety information. *Am Ind Hyg Assoc J* 1993;**54**:683–96.
8 Thirion B, Darmoni SJ, Moore N. Costs of Medline and CD-ROM searching. *Lancet* 1992;**340**:308.