# Search strings for the study of putative occupational determinants of disease

Stefano Mattioli,[1] Francesca Zanardi,[1] Alberto Baldasseroni,[2] Frederieke Schaafsma,[3] Robin MT Cooke,[1] Gianpiero Mancini,[4] Mauro Fierro,[1] Chiara Santangelo,[1] Andrea Farioli,[1] Serenella Fucksia,[1] Stefania Curti,[1] Francesco S Violante,[1] Jos Verbeek[5]

► A supplementary table, box and appendix are published online only. To view these files please visit the journal online (http://oem.bmj.com).

[1]Section of Occupational Medicine, Department of Internal Medicine, Geriatrics and Nephrology, University of Bologna, Italy
[2]Tuscany Regional Centre for Occupational Injuries and Diseases (CeRIMP), Florence, Italy
[3]Coronel Institute of Occupational Health, Academic Medical Centre, Amsterdam, the Netherlands
[4]Department of Prevention, Azienda USL di Ravenna, Ravenna, Italy
[5]Finnish Institute of Occupational Health, Centre of Good Practices, Knowledge Transfer Team, Cochrane Occupational Health Field, Kuopio, Finland

**Correspondence to**
Francesco S Violante, Unità Operativa di Medicina del Lavoro, Dipartimento di Medicina Interna, dell'Invecchiamento e Malattie Nefrologiche, Università di Bologna, Policlinico S. Orsola-Malpighi, Via Pelagio Palagi 9, I - 40138 Bologna, Italy;
francesco.violante@unibo.it

Accepted 9 September 2009

## ABSTRACT

**Objective** To identify efficient PubMed search strategies to retrieve articles regarding putative occupational determinants of conditions not generally considered to be work related.

**Methods** Based on MeSH definitions and expert knowledge, we selected as candidate search terms the four MeSH terms describing 'occupational disease', 'occupational exposure', 'occupational health' and 'occupational medicine' (DEHM) alongside 22 other promising terms. We first explored overlaps between the candidate terms in PubMed. Using random samples of abstracts retrieved by each term, we estimated the proportions of articles containing potentially pertinent information regarding occupational aetiology in order to formulate two search strategies (one more 'specific', one more 'sensitive'). We applied these strategies to retrieve information on the possible occupational aetiology of meningioma, pancreatitis and atrial fibrillation.

**Results** Only 20.3% of abstracts were retrieved by more than one DEHM term. The more 'specific' search string was based on the combination of terms that yielded the highest proportion (40%) of potentially pertinent abstracts. The more 'sensitive' string was based on the use of broader search fields and additional coverage provided by other search terms under study. Using the specific string, the numbers of abstracts needed to read to find one potentially pertinent article were 1.2 for meningioma, 1.9 for pancreatitis and 1.8 for atrial fibrillation. Using the sensitive strategy, the numbers needed to read were 4.4 for meningioma, 8.9 for pancreatitis and 10.5 for atrial fibrillation.

**Conclusions** The proposed strings could help health care professionals explore putative occupational aetiology for diseases that are not generally thought to be work related.

### What this paper adds

- Well defined PubMed search strategies can provide efficient and effective tools to help address evidence-based questions.
- Such strategies are currently available for diseases that have been widely studied from the occupational standpoint, but not for exploration of potential 'new' occupational diseases.
- The two search strings presented in this article provide readily applicable 'copy and paste' tools for prompt identification of pertinent literature available in PubMed.
- The more specific of the two strings is capable of retrieving a high proportion of potentially pertinent abstracts related to diseases that have been little studied from the occupational standpoint; the more sensitive string may be used for more extensive research or to retrieve information regarding diseases that have received even less attention.
- These search strings could be useful to many different types of health professional, ranging from primary care physicians to specialised librarians, in contexts ranging from evidence-based patient evaluation to original research.

## INTRODUCTION

Some diseases can have rather obscure occupational determinants (eg, increased susceptibility to infectious pneumonia in workers exposed to metal fumes[1]). Well defined PubMed search strategies[2] can provide efficient and effective tools for answering evidence-based questions in the field of occupational health,[3] and exploring the possible work-related aetiology of given diseases.[4] Such knowledge can provide an important basis for application of evidence-based medicine and evidence-based prevention in occupational health.

The controlled, hierarchical vocabulary of Medical Subject Heading (MeSH) terms provides a consistent way of retrieving articles deemed pertinent to specific areas of medical interest, and therefore provides a powerful tool for focussing PubMed searches. As health professionals now commonly use bibliographic searches via PubMed to get answers to practice-related questions, rational use of MeSH terms is becoming increasingly important.[5] Due to terminological overlaps in the MeSH vocabulary, variations inevitably occur when MeSH terms are systematically assigned to articles during manual indexing at the US National Library of Medicine. Thus, specific PubMed search strategies need to be developed and evaluated for particular areas of investigation. A set of rational PubMed search strategies has been developed for study of the work-related origins of various classes of diseases that have attracted extensive occupational health research.[4] A randomised controlled trial showed that these strategies could effectively enhance adequate selection of search terms, satisfactory solutions to case problems, and user

**Box 1 Proposed PubMed search strategies for identifying potentially pertinent articles**

1. More specific search strategy:
*(occupational diseases [MH] OR occupational exposure [MH] OR occupational medicine [MH] OR occupational risk [TW] OR occupational hazard [TW] OR (industry [MeSH Terms] mortality [SH]) OR occupational group* [TW] OR work-related OR occupational air pollutants [MH] OR working environment [TW])* AND name(s)-of-the-disease.
2. More sensitive search strategy:
*(occupational diseases [MH] OR occupational exposure [MH] OR occupational exposure* [TW] OR "occupational health" OR "occupational medicine" OR work-related OR working environment [TW] OR at work [TW] OR work environment [TW] OR occupations [MH] OR work [MH] OR workplace* [TW] OR workload OR occupation* OR worke* OR work place* [TW] OR work site* [TW] OR job* [TW] OR occupational groups [MH] OR employment OR worksite* OR industry)* AND name(s)-of-the-disease.

Usage notes

1. It is possible to 'copy and paste' each of the two strings into PubMed from a .doc file. Alternatively, the strings can be evoked in PubMed by entering the following shortened URLs in the browser address box:
2. http://tinyurl.com/mattioli-et-al-specific for the 'more specific' string and http://tinyurl.com/mattioli-et-al-sensitive for the 'more sensitive' string.
3. The name-of-the-disease should be entered without any search tag. For diseases that have more than one name, the various 'names-of-the-disease' should be entered in brackets, connected by the OR operator: for example … AND (epicondylitis OR tennis elbow).

Technical notes

When a MeSH term contains two words, PubMed's automatic query translation currently comprises a search for the entire MeSH term plus All Fields searches for the two words (eg, 'occupational medicine'[MeSH Terms] OR ('occupational'[All Fields] AND 'medicine'[All Fields]) OR 'occupational medicine'[All Fields]).

In the presence of an 'embedded' MeSH Term (eg, as in the case of *occupational risk*, where *risk* is also a MeSH Term) PubMed automatically also searches for the MeSH Term of the single word.

In the more sensitive string, both 'occupational health' and 'occupational medicine' are entered in an All Fields format which evokes all the abstracts retrieved when these two search terms were entered in the [MeSH Terms] OR [Text Word] field.

satisfaction.[6] However, effective and efficient PubMed search strategies need to be defined for conditions not commonly considered to be work-related diseases but which could plausibly have occupational determinants.

The aim of this study was to develop efficient PubMed search strings to help assess the existence of putative occupational determinants of conditions that are not generally considered to be work related.

## METHODS

### Rationale and study design

Since it is not feasible to study all possible search terms, in a preliminary phase we identified sets of MeSH and non-MeSH terms that seem especially pertinent to occupational determinants of diseases. Our strategy was to select the broadest general descriptors available in the medical MeSH vocabulary, alongside other MeSH/non-MeSH terms which might conceivably help retrieve further pertinent literature or refine search strategies. As a basis for formulation of specific search strategies, we assessed the retrieval characteristics of selected search terms when used separately or in combination. We first explored their 'coverage' within PubMed in terms of numbers of articles identified by each term. We also explored overlaps between terms (numbers of articles shared by different terms) so as to get an indication of their mutual exclusiveness. For each of the search terms we estimated proportions of articles carrying English language abstracts that could be considered potentially pertinent to the field of occupational aetiology. Based on these findings, we formulated two search strategies (one more 'specific', one more 'sensitive') designed for use in different circumstances. Finally, we assessed 'number needed to read' (NNR) values by applying these two strategies to three diseases that are not generally thought to be work related.

All bibliometric data were generated with the date limit function set to call up articles added to PubMed by 14 February 2008.

### Selection of terms to be tested

Using the Medline MeSH database, we first considered work-related MeSH terms, such as those evoked by the terms *'occupational'* (n=57),*'work'* (n=15), *'job'* (n=12), etc (see online Appendix 1), along with their various subheadings. We decided to focus first on a group of four MeSH terms with especially broad definitions which could be pertinent to occupational aetiology (*occupational diseases*, *occupational exposure*, *occupational health*, *occupational medicine*). We reasoned that this group of four MeSH descriptors covering disease, exposure, health and medicine (DEHM group) appears to target four broad areas of relevance to occupational aetiology. Based on the particular definitions and retrieval characteristics of other work-related MeSH terms (see online Appendix 1) and on preliminary studies (not shown), we also decided to evaluate eight other work-related MeSH terms (with/without subheadings) that suggested a potential to expand or modulate search strategies: namely, *employment*; *industry*; *occupations*; *occupational air pollutants*; *occupational groups*; *work*; *workload*; *workplace*. Regarding search terms that fall outside the MeSH vocabulary, choices were based on findings from a single available study on PubMed searches regarding occupational aetiology,[4] review of MeSH entry terms, the authors' experience and brainstorming (all in conjunction with preliminary, sample PubMed searches). After extensive exploration (not shown), we eventually selected 14 items: *at work[Text Word]*; *industrial hygiene[Text Word]*; *job*[Text Word]*; *occupation**; *occupational hazard[Text Word]*; *occupational risk[Text Word]*; *worke**; *work environment[Text Word]*; *work-related*; *working environment[Text Word]*; *workplace**; *work place*[Text Word]*; *worksite*[Text Word]*; *work site*[Text Word]*. Of note, we chose to incorporate the [Text Word] search tag by default into 10 of these terms in order to avoid undesired automatic term mapping (see technical note in box 1). When not otherwise stated, search terms were entered 'untagged' to take advantage of PubMed's automatic term mapping algorithms.

To get indications of the numbers of articles identified by each of the 26 selected search terms (ie, the DEHM descriptors plus the eight other MeSH terms and the 14 non-MeSH items), we also used PubMed limits functions to calculate proportions of Medline articles in selected languages, as well as the proportion of articles in any language with an available English-language abstract (see online table 1).

### Survey of search term overlaps

For each DEHM descriptor (*occupational diseases*, *occupational exposure*, *occupational health* and *occupational medicine*), we recorded the number of articles with available abstracts identified in PubMed using each of the following search fields: (1) [MeSH Terms]; (2) [Text Word] NOT [MeSH Terms]; (3) [All Fields] NOT ([MeSH Terms] OR [Text Word]). We used a similar approach to assess the other selected MeSH terms (*employment*; *industry*; *occupations*; *occupational air pollutants*; *occupational groups*; *work*; *workload*; *workplace)* and non-MeSH search terms (*at work[Text Word]*; *industrial hygiene[Text Word]*; *job*[Text Word]*; *occupation**; *occupational hazard[Text Word]*; *occupational risk[Text Word]*; *worke**; *work environment[Text Word]*; *work-related*; *working environment[Text Word]*; *workplace**; *work place*[Text Word]*; *work-site*[Text Word]*; *work site*[Text Word]*). Of note, to avoid semantically inappropriate automatic term mapping we entered the MeSH descriptor *work* only in the [MeSH Terms] field.

We then assessed retrieval overlaps (and omissions) between each of the four DEHM terms. To do this, we crossed the DEHM terms (two, three or four at a time, using Boolean operators) within the '[MeSH Terms] OR [Text Word]' search field in such a way as to record numbers of articles identified for each of their possible combinations. Since availability of an English language abstract can be of practical importance when assessing the potential relevance of an article, we decided also to introduce the limit 'Abstracts'. Finally, we used Boolean operators to assess overlaps between each of the 16 non-DEHM search items (entered without additional tags, other than those specified above) and the entire DEHM group (entered using the search field '[MeSH Terms] OR [Text Word]').

### Estimating proportions of pertinent articles

Estimates were based on samples of 100 articles with available abstracts which were randomly extracted on entering the search terms under study in PubMed (using defined search field tags) in conjunction with the 'Abstract' limit function. For each (tagged/untagged) search term under investigation, the random sample was obtained by setting the PubMed 'show' function in such a way as to obtain a number of pages approximately corresponding to a multiple of 100: we then extracted abstracts for 'top-of-the-page' articles (after regularly skipping appropriate numbers of pages). The pertinence of each article was assessed by two occupational physicians (GM, MF) who independently examined each abstract and expressed a binary judgement based on presence of information regarding evidence or hypotheses (irrespective of study design) regarding occupational determinants of disease. (Regarding interobserver variability, in a preliminary assessment of 100 abstracts, the two observers achieved a $\kappa$ value of 0.79 (SE 0.099), corresponding to 'good' agreement.[7]) In cases of disagreement, pertinence was adjudicated by a third physician (SM).

We first assessed the pertinence of the entire group of DEHM terms (entered with the OR operator) in different search tag combinations: (1) [MeSH Terms]; (2) [Text Word] NOT [MeSH Terms]; (3) [All Fields] NOT ([MeSH Terms] OR [Text Word]). These search fields were selected so as to provide indications of the incremental yield of pertinent articles provided by the Text Word and All Fields tags. Additionally, we estimated the proportions of potentially pertinent abstracts retrieved by each of the DEHM terms when entered as [MeSH Terms]. We then assessed the possible incremental pertinence of each of the 22 other search items (entered as listed above), while excluding the entire DEHM group (entered as '[MeSH Terms] OR [Text Word]').

Based on these findings, we devised two distinct search strategies to be proposed for routine use: one designed to be more specific ('first string') and one rather more sensitive ('second string'). Of note, selection of the cut-off used to define the more specific string (>40% of pertinent articles) was loosely based on the proportion of potentially pertinent articles retrieved by the entire DEHM group, which also corresponded to a NNR value of <2.5.

### Assessment of NNR values

We evaluated the number of abstracts needed to read to identify one potentially pertinent article in the context of three different pathologies not generally thought to be work related: namely, 'pancreatitis', 'atrial fibrillation' and 'meningioma'. For each pathology, we retrieved all the abstracts evoked by each of the two candidate search strategies. The same team of readers (ie, GM, MF and SM) assessed the pertinence of each abstract using the rating criteria described above. We then calculated the NNR values for each string.[8] We also calculated NNR values for two other proposed search strategies: (1) the string developed by Schaafsma *et al* for use by physicians looking for literature regarding diseases that have attracted more widespread study of possible occupational aetiology, that is *(occupational risk OR occupational disease) AND name(s)-of-the-disease*[4]; (2) a string developed by the Cochrane Occupational Health Field for locating occupational health studies referring to work, that is *(occupat* OR worker*) AND name(s)-of-the-disease*.[3] Finally, we explored the effects of combining our first (more specific) string with the two (narrow/broad) aetiology search filters provided by PubMed for clinical queries regarding specific clinical study categories.[9 10]

## RESULTS

### Numbers of articles identified and their overlaps

#### Role of DEHM terms

Table 1 reports the numbers of articles (and abstracts) identified by each of the four DEHM MeSH descriptors (*occupational diseases*, *occupational exposure*, *occupational health*, *occupational medicine*) using different combinations of search tags. Entering the entire DEHM group in a rather comprehensive search field ('[MeSH Terms] OR [All Fields]') identified 170 316 articles (78 053 abstracts) from PubMed, representing ~1% (~1% abstracts) of all 17 884 312 PubMed articles (9 542 808 abstracts). Of note, *'occupational diseases'* identified the highest number of articles (two to four times the numbers identified by each of the other three terms).

The [MeSH Terms] search tag was more productive than the [Text Word] tag when used with *occupational diseases* and *occupational exposure*, but not with *occupational health* or *occupational medicine*. Incorporation of the [All Fields] tag identified substantial numbers of additional articles for *occupational health* (7228 more abstracts) and *occupational medicine* (3976 more abstracts), due to frequent appearance of these search terms in the affiliations or journal names but not in the title or main text of the article.

Figure 1 reports the relative coverage and overlaps provided by each of the DEHM terms when entered with the '[MeSH Terms] OR [Text Word]' search field and using the limit 'Abstracts'. Overlaps between two or more of the DEHM terms were observed for only one fifth (20.3%, 14 500/71 264) of the abstracts identified. Remarkably, only 60 (0.08%) abstracts were retrieved by all four DEHM terms. The largest single overlap was between *occupational exposure* and *occupational diseases* (about a quarter of the articles identified by *occupational exposure* were

**Table 1** Numbers of articles (abstracts) identified by each of the DEHM terms* when using different PubMed search fields

| | DEHM terms | | | |
|---|---|---|---|---|
| | **Occupational diseases** | **Occupational exposure** | **Occupational health** | **Occupational medicine** |
| [MeSH Terms] | 92383 (39331) | 36193 (25358) | 17844 (7976) | 20862 (3128) |
| [Text Word] | 65162 (29299) | 33511 (26609) | 32148 (13745) | 22101 (4042) |
| [All Fields] | 65441 (29556) | 33511 (26609) | 40182 (20973) | 26422 (8018) |
| [MeSH Terms] OR [Text Word] | 92994 (39560) | 40577 (29209) | 32148 (13745) | 22101 (4092) |
| [All Fields] NOT ([MeSH Terms] OR [Text Word])† | 241 (223) | 0 (0) | 8034 (7228) | 4321 (3976) |
| [MeSH Terms] OR [Text Word] OR [All Fields]‡ | 93235 (39783) | 40577 (29209) | 40182 (20973) | 26422 (8018) |

*DEHM terms: *occupational diseases*, *occupational exposure*, *occupational health*, *occupational medicine*.
†Search field revealing numbers of articles in which the search terms appear only in the 'affiliations' or 'journal names'.
‡Most comprehensive search field.

also netted by *occupational diseases*). Of note, similar results were obtained when the searches were run without the 'Abstracts' limit (data not shown).

### *Role of non-DEHM search terms*

We evaluated the ability of each of the 22 non-DEHM search terms to identify abstracts not caught by the DEHM group (table 2). Overall, the non-DEHM search terms netted 802 891 articles (402 245 containing abstracts), representing about 4% of all articles listed in PubMed. Of these, 648 707 (80.8%) were not caught by the DEHM group, including 328 913 with available abstracts. The potential incremental contribution of non-DEHM search terms can also be discerned by considering numbers of articles (with/without abstracts) identified in different languages (see online table 1). In particular, a remarkably high proportion (23%) of all articles bearing the DEHM MeSH term *occupational medicine* are in Russian (with German accounting for a further 10%). On the other hand, the DEHM MeSH term *occupational health* is relatively little represented in languages other than English. Such variability in assigning MeSH terms provides a further rationale for attempting a broader search strategy.

## Pertinence estimates

We first assessed proportions of articles potentially pertinent to occupational aetiology based on randomly extracted abstracts when entering the entire DEHM group (ie, *occupational diseases [MH] OR occupational diseases [TW] OR occupational exposure [MH] OR occupational exposure [TW] OR occupational health [MH] OR occupational health [TW] OR occupational medicine [MH] OR occupational medicine [TW]*). Use of different search tag combinations (chosen to evaluate use of the [MeSH Terms] field and possible incremental contributions of other fields) provided the following results: 48% potentially pertinent abstracts using [MeSH Terms]; 17% using '[Text Word] NOT [MeSH Terms]'; 15% using '[All Fields] NOT ([MeSH Terms] OR [Text Word]'). These figures suggest that 48% of the abstracts retrieved by the [MeSH Terms] field may be pertinent, along with 17% of those additionally retrieved by incorporating the [Text Word] tag, and about 15% of those incrementally retrieved by additionally incorporating [All Fields]. Figure 2 illustrates these findings in relation to the total numbers of abstracts retrieved by the DEHM group using these three search field combinations.

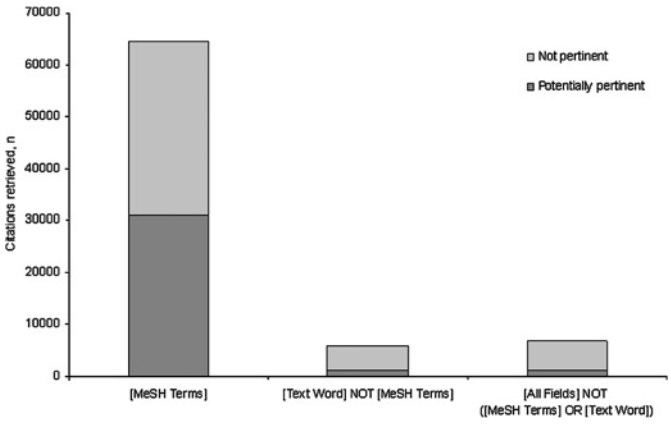


**Figure 1** Proportional Venn diagram representing numbers of abstracts retrieved from PubMed using different possible combinations of the DEHM terms (*occupational diseases*, *occupational exposure*, *occupational health* and *occupational medicine*) using [MeSH Terms] OR [Text Word]).

We then looked at numbers of potentially pertinent abstracts retrieved by each of the DEHM terms when entered as [MeSH Terms]: the estimated proportions were 62% for occupational diseases, 58% for occupational exposure, 30% for occupational medicine and 27% for occupational health. Regarding the incremental contributions of non-DEHM terms, table 2 also reports proportions of randomly retrieved abstracts that were deemed potentially pertinent when each of these terms was entered using the search field '[MeSH Terms] OR [Text Word]' after exclusion of articles retrieved by the entire DEHM group.

## Formulation of search strings

The two proposed search strings are presented in box 1. The more specific search strategy ('first string') included those search terms which were estimated to retrieve >40% of pertinent articles (corresponding to a NNR value <2.5). Additionally, we decided to include *occupational medicine[MeSH Terms]*, based on the observation that until the mid-1980s this MeSH term was ascribed to many potentially pertinent articles (data not shown)—a relevant consideration when exploring the aetiology of diseases that have been little studied from the occupational standpoint.

To try to make the strategy more sensitive ('second string'), we (1) broadened the search fields for each of the DEHM descriptors to [MeSH Terms] OR [Text Word] OR [All Fields] (table 1) and (2) took advantage of the additional coverage provided by the other search terms under study (table 2), except for *occupational risk [TW]*, *occupational hazard [TW]*, *occupational group*[TW]* and *occupational air pollutants [MH]* which did not identify any incremental articles (beyond those already evoked by the remaining terms).

**Table 2** Incremental contribution of non-DEHM search terms: numbers of abstracts not identified by the DEHM group† (and estimates of numbers potentially pertinent to occupational aetiology)

| PubMed query | PubMed query translation (automatic term mapping) | Proportion (%, *n*/*N*) of the abstracts retrieved by the search term (*N*) which were not retrieved by the DEHM group (*n*)‡ | Estimated proportion of potentially pertinent additional abstracts (%)§ | Estimated absolute numbers of potentially pertinent additional abstracts¶ |
|---|---|---|---|---|
| occupational hazard[Text Word] | occupational hazard[Text Word] | 32% (172/531) | 77 | 132 |
| industry[MeSH Terms] mortality [Subheading] | 'industry'[MeSH Terms] AND 'mortality'[Subheading] | 32% (508/1585) | 74 | 376 |
| occupational risk[Text Word] | occupational risk[Text Word] | 21% (258/1207) | 73 | 188 |
| occupational group*[Text Word] | *From* occupational group[Text Word] *…to…* occupational groups[Text Word] | 46% (764/1670) | 52 | 397 |
| work-related | work-related[All Fields] | 42% (2266/5407) | 46 | 1042 |
| occupational air pollutants[MeSH Terms] | 'air pollutants, occupational'[MeSH Terms] | 33% (2316/6942) | 45 | 1042 |
| working environment[Text Word] | working environment[Text Word] | 50% (647/1308) | 41 | 265 |
| at work[Text Word] | at work[Text Word] | 68% (4521/6642) | 36 | 1628 |
| work environment[Text Word] | work environment[Text Word] | 55% (1547/2834) | 33 | 511 |
| occupations[MeSH Terms] | occupations[MeSH Terms] | 78% (8247/10521) | 32 | 2639 |
| work[MeSH Terms] | work[MeSH Terms] | 76% (3519/4642) | 31 | 1091 |
| workplace* | *From* workplace[All Fields] *…to…* OR workplacetraumacenter[All Fields] | 50% (7621/15315) | 28 | 2134 |
| Workload | Workload[MeSH Terms] OR workload[All Fields] | 90% (12351/13728) | 27 | 3335 |
| occupation* | *From* occupation[All Fields] *…to…* occupationwise[All Fields] | 46% (55339/119619) | 25 | 13835 |
| worke* | *From* worke[All Fields] *…to…* workey[All Fields] | 66% (59185/90267) | 20 | 11837 |
| work place*[Text Word] | work place[Text Word] OR work placement[Text Word] OR work placements[Text Word] OR work places [Text Word] | 58% (685/1196) | 19 | 130 |
| work site*[Text Word] | work site[Text Word] OR work sites[Text Word] OR work sitewide[Text Word] | 45% (340/758) | 15 | 51 |
| job*[Text Word] | *From* job[Text Word] *…to…* jobwatch[Text Word] | 79% (24348/30715) | 14 | 3409 |
| occupational groups[MeSH Terms] | 'occupational groups'[MeSH Terms] | 94% (108099/115419) | 9 | 9729 |
| employment | 'employment'[MeSH Terms] OR 'employment'[All Fields] | 83% (32187/38636) | 9 | 2897 |
| worksite* | *From* worksite[All Fields] *…to…* worksitewide[All Fields] | 39% (575/1469) | 6 | 34 |
| industry | 'industry'[MeSH Terms] OR 'industry'[All Fields] | 86% (91082/105760) | 4 | 3643 |

In occupation* and worke*, the asterisk represents the PubMed truncation symbol. MeSH, Medical Subject Heading.
†DEHM group: (*occupational diseases* OR *occupational exposure* OR *occupational medicine* OR *occupational health*) entered using the search field [MeSH Terms OR Text Word].
‡Calculated as: ((total number identified by the non-DEHM term)−(number also identified by the DEHM group))/(total number identified by the non-DEHM term).
§Estimates were based on reviews of 100 randomly sampled abstracts.
¶Calculated by multiplying the number of abstracts additionally identified (ie, *n* in column 2) by the estimated proportion of potentially pertinent additional abstracts (column 3).

### Assessment of proposed search strategies

We assessed the characteristics of the two proposed search strategies (alongside two other strategies proposed elsewhere[3 4]) in three pathologies not commonly thought to be work related: namely, 'meningioma', 'atrial fibrillation' and 'pancreatitis'. Table 3 reports the numbers of abstracts retrieved by each strategy, together with the proportions of retrieved abstracts that were deemed pertinent and their NNR values. For each pathology, the NNR values were lowest for the 'more specific' strategy ('first string') and highest for the 'more sensitive' strategy ('second string'). Furthermore, the 'second string' invariably retrieved the highest absolute number of pertinent abstracts. The two strategies proposed elsewhere[3 4] appeared to display intermediate characteristics in terms of both their NNR values and the absolute numbers of pertinent articles retrieved. Finally, we found that entering our 'first string' in conjunction with the narrow/broad aetiology search filters provided by PubMed for clinical queries regarding specific clinical study

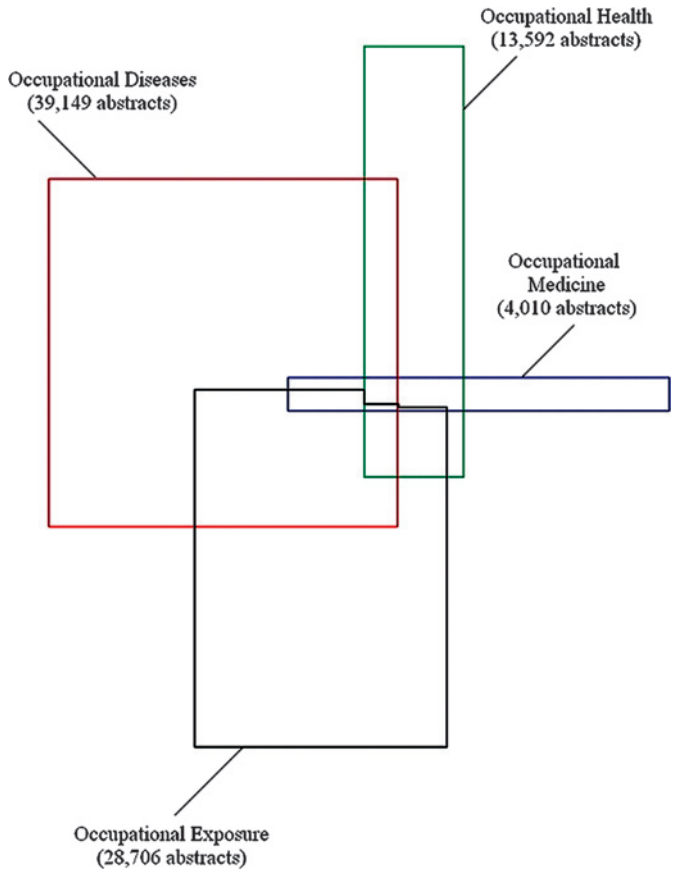


**Figure 2** Bar chart showing the limited incremental contribution of the [Text Word] and [All Fields] search fields* for retrieval of potentially relevant articles when entering the entire DEHM group of MeSH terms (ie, *occupational diseases* OR *occupational exposure* OR *occupational health* OR *occupational medicine*) using different search fields: (a) [MeSH Terms]; (b) [Text Word] NOT [MeSH Terms]; (c) [All Fields] NOT ([MeSH Terms] OR [Text Word]). *To search a MeSH term using the [All Fields] option, it is necessary to enter the term in inverted commas (eg, *'occupational diseases'*).

categories[9] reduced the numbers of abstracts identified without improving any of the NNR values.

## DISCUSSION

This bibliometric study proposes two readily applicable PubMed search strings (one more specific, one more sensitive; box 1) for use by health professionals when investigating putative occupational determinants of medical conditions that are not generally classified as work related. These strings are intended to complement previously proposed and tested strings designed for evaluation of occupational aetiology in more widely studied diseases.[4]

Initially, we decided to take advantage of the relatively exclusive search characteristics of the four MeSH descriptors most broadly dedicated to occupation or work (*occupational diseases*, *occupational exposure*, *occupational health*, *occupational medicine*). Perhaps due to their explicit focus on different broad areas of relevance to occupational aetiology (ie, disease, exposure, health and medicine; termed DEHM by us), we found that only one fifth of the abstracts netted by any one of these four terms could also be retrieved by one of the other DEHM terms (see figure 1). We eventually included *occupational health* only in the more sensitive string (due to its relatively low specificity).

**Table 3** Application of search strategies to three pathologies: numbers of citations retrieved, proportions of potentially pertinent articles and overall NNR values

| PubMed query | Meningioma (n=9460) | | | Pancreatitis (n=24497) | | | Atrial fibrillation (n=20313) | | | Overall (n=54247) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Retrieved, n | Pertinent, n (%) | NNR | Retrieved, n | Pertinent, n (%) | NNR | Retrieved, n | Pertinent, n (%) | NNR | Retrieved, n | Pertinent, n (%) | NNR |
| 'First string' (more specific strategy) | 20 | 16 (80) | 1.2 | 19 | 10 (53) | 1.9 | 20 | 11 (55) | 1.8 | 59 | 37 (63) | 1.6 |
| 'First string'+narrow PubMed clinical query for aetiology[9] | 5 | 4 (80) | 1.2 | 5 | 1 (20) | 5.0 | 2 | 1 (50) | 2.0 | 12 | 6 (50) | 2.0 |
| 'First string'+broad PubMed clinical query for aetiology [9] | 17 | 14 (82) | 1.2 | 14 | 6 (43) | 2.3 | 6 | 2 (33) | 3.0 | 37 | 22 (59) | 1.7 |
| 'Second string' (more sensitive strategy) | 80 | 18 (23) | 4.4 | 195 | 22 (11) | 8.9 | 232 | 22 (9) | 10.5 | 507 | 62 (12) | 8.2 |
| 'Second string' NOT 'first string' (incremental contribution of the second string) | 60 | 2 (3) | 30.0 | 176 | 12 (7) | 14.7 | 212 | 11 (5) | 19.3 | 448 | 25 (6) | 17.9 |
| String proposed by Schaafsma et al[4]† | 20 | 13 (65) | 1.5 | 36 | 13 (36) | 2.8 | 18 | 7 (39) | 2.6 | 74 | 33 (45) | 2.2 |
| String used by the Cochrane Occupational Health Field[3]‡ | 51 | 17 (33) | 3.0 | 108 | 18 (17) | 6.0 | 60 | 15 (25) | 4.0 | 220 | 50 (23) | 4.4 |

†(Occupational risk OR occupational diseases) AND name-of-the-disease.
‡(Occupat* OR worker*) AND name-of-the-disease.
NNR, number needed to read value.

Furthermore, another MeSH descriptor, *occupational air pollutants*, turned out to play a useful role in the specific string.

Consideration of the NNR suggests that the more specific string is likely to provide a much more attractive way of addressing many questions encountered in routine practice. In each of the three diseases we looked at, as many as 63% of the abstracts retrieved by the specific string appeared to be potentially pertinent. By contrast, it would be necessary on average to scan four or five abstracts concerning meningioma retrieved by the sensitive string to identify one potentially pertinent paper (and for atrial fibrillation and pancreatitis the NNR was as high as 9 or 10). These findings suggest that the first (more specific) string may provide an efficient frontline approach for healthcare professionals who need to explore the putative occupational aetiology of little studied diseases in practice-based situations ranging from primary care to medicolegal issues or insurance claims. This concept is reinforced by comparison of the numbers of pertinent abstracts (with their corresponding NNR) obtained by this string and by the string proposed by Schaafsma *et al*[4] for use with diseases that have been more widely studied from an occupational standpoint (overall NNR for the three diseases considered, 1.6 vs 2.2; table 3). However, we would not recommend the string for more widely studied diseases such as carpal tunnel syndrome, where our more specific, but still relatively lengthy, string retrieved only a few more abstracts (583 vs 562) than the much more compact string proposed by Schaafsma *et al* (data not shown). Of note, one conceivable way of reducing the number of abstracts identified by the specific string without greatly raising the NNR value might be to incorporate one of the two aetiology search filters provided by PubMed for clinical queries.[9]

The second (more sensitive) string developed in the present study could be adopted either to assess diseases which elicit only a few articles or to explore a little studied disease in more depth. We also suspect that the sensitive string may provide a useful point of departure for more exhaustive investigations, such as systematic reviews of the literature conducted for research or medicolegal purposes. Comparison of the numbers of pertinent abstracts retrieved by this string and by the one used in the Cochrane Occupational Health Field[3] for the three tested pathologies (table 3) suggests that our 'second string' may be more sensitive. We also tried applying the 'second string' to retinal detachment (data not shown), a disease that has been very little studied from the standpoint of occupational aetiology. In this challenging context, the sensitive string retrieved a total of 125 articles (80 with abstracts), only four (2%) of which appeared to be potentially pertinent. Despite the small haul of pertinent articles, research experience[11] suggests that the string could provide an efficient tool for initial research, with the potential to save time by rapidly retrieving most of the available articles (although we do know it missed at least one pertinent paper[12]). Interestingly, when we searched for abstracts published in journals listed in the Occupational Medicine subset of the NCBI Journals database (online box 1) that were not retrieved by the 'second string' (data not shown), only one such abstract turned up for the three diseases under study (and it did not appear to be pertinent).

### Study limitations

The practical decision to base the assessments of pertinence on articles with available (English language) abstracts may have led to some selection bias due to exclusion of certain article types, such as letters, which could contain relevant information. However, a supplemental analysis based on information contained in titles suggested that this factor would not have constituted a major bias (data not shown). The assessments could not take into account relevant information reported in the main body of the article but not in abstracts, the quality of which can vary a lot—especially in the absence of widespread provision of more informative abstracts.[13] Furthermore, we did not attempt to evaluate the quality of the individual studies. Although power calculations would have enhanced the precision of the estimated proportions of pertinent articles, this factor would not have substantially affected the main study objective (identification of efficient search strings). Since no gold standard instrument exists for retrieval of pertinent articles, we were unable to evaluate sensitivity and specificity values for the two proposed search strings (although the NNR values do give some indication of specificity). It could be argued that our selection of non-MeSH search terms was to some extent arbitrary. However, the ability of the more sensitive (second) string to retrieve most of the available pertinent abstracts for a range of diseases (see above) suggests that this a priori limitation did not greatly impact on the end product. We decided not to attempt to incorporate the Occupational Medicine subset of the Journals Database because of the unwieldy length of the resulting search string (online box 1) in the current absence of a dedicated PubMed 'Journals Group' search tag (equivalent to the subset tag [st] in the Journals Database). In any case, it should be underlined that this study was restricted to PubMed: systematic reviews of the literature would require additional bibliographic searches using other relevant search engines, such as Embase.[14] Changes in research and reporting practices (eg, choice of key words) over time[10] will inevitably affect the retrievability of future literature. For instance, implementation of STROBE guidelines[15] could (hopefully) improve the reporting quality of titles and abstracts of epidemiological studies, thereby facilitating identification of pertinent articles.

## CONCLUSIONS

In conclusion, box 1 reports two proposed PubMed search strings—one more specific, one more sensitive—which may be used for rapid (or more lengthy) explorations of evidence regarding the existence of possible occupational determinants of a disease that is not generally thought to be work related. Either string can be pasted into the PubMed search box alongside the name(s)-of-the-disease (see box 1). We recommend trying the more specific string first and then, if necessary, the second string. About half the articles retrieved by the first string are likely to be potentially pertinent to occupational aetiology in general. We think that such retrieval characteristics could make this straightforward tool useful in a variety of health practice situations. Field tests are required to assess the effectiveness of applying these strategies in the real world.

**Acknowledgements** We are particularly grateful to Eva Buiatti for her highly valued encouragement. We would also like to thank Melvin Piro for generating the proportional Venn diagram. Claudio Giampaoletti assisted in the preliminary phases of the study. Giusi Vasta and Alessandro Catanese helped perform PubMed searches.

**Funding** INAIL (Istituto Nazionale per l'Assicurazione contro gli Infortuni sul Lavoro) Direzione Regionale Emilia-Romagna, Bologna, Italy; ISPESL (Istituto Superiore per la Prevenzione e la Sicurezza del Lavoro), Rome, Italy; Regione Emilia-Romagna (Emilia-Romagna Regional Administration), Bologna, Italy; and the University of Bologna provided funding for this study.

**Competing interests** None.

**Provenance and peer review** Not commissioned; externally peer reviewed.

## REFERENCES


1. **Palmer KT,** Poole J, Ayres JG, *et al*. Exposure to metal fume and infectious pneumonia. *Am J Epidemiol* 2003;**157**:227—33.
2. **Gehanno JF,** Thirion B. How to select publications on occupational health: the usefulness of Medline and the impact factor. *Occup Environ Med* 2000;**57**:706—9.
3. **Verbeek J,** Salmi J, Pasternack I, *et al*. A search strategy for occupational health intervention studies. *Occup Environ Med* 2005;**62**:682—7.
4. **Schaafsma F,** Hulshof C, Verbeek J, *et al*. Developing search strategies in Medline on the occupational origin of diseases. *Am J Ind Med* 2006;**49**:127—37.
5. **Lowe HJ,** Barnett GO. Understanding and using the medical subject headings (MeSH) vocabulary to perform literature searches. *JAMA* 1994;**271**:1103—8.
6. **Schaafsma F,** Hulshof C, de Boer A, *et al*. Effectiveness and efficiency of a literature search strategy to answer questions on the etiology of occupational diseases: a controlled trial. *Int Arch Occup Environ Health* 2007;**80**:239—47.
7. **Altman DG.** ed. Some common problems in medical research. In: *Practical statistics for medical research*. London: Chapman and Hall, 1991:403—9.
8. **Bachmann LM,** Coray R, Estermann P, *et al*. Identifying diagnostic studies in MEDLINE: reducing the number needed to read. *J Am Med Inform Assoc* 2002;**9**:653—8.
9. PubMed Clinical queries. http://www.ncbi.nlm.nih.gov/entrez/query/static/clinical.shtml (accessed 17 Mar 2009).
10. **Wilczynski NL,** Haynes RB, for the Hedges Team. Developing optimal search strategies for detecting clinically sound causation studies in MEDLINE. *AMIA Annu Symp Proc* 2003:719—23.
11. **Mattioli S,** De Fazio R, Buiatti E, *et al*. Physical exertion (lifting) and retinal detachment among people with myopia. *Epidemiology* 2008;**19**:868—71.
12. **Pivovarov NN,** Malakhova LA, Bagdasarova TA, *et al*. Role of weight lifting in the development of retinal detachment. *Vestn Oftalmol* 1977;**6**:50—3.
13. **Gehanno JF,** Paris C, Thirion B, *et al*. Assessment of bibliographic databases performance in information retrieval for occupational and environmental toxicology. *Occup Environ Med* 1998;**55**:562—6.
14. **Haynes RB,** Mulrow CD, Huth EJ, *et al*. More informative abstracts revisited. *Ann Intern Med* 1990;**113**:69—76.
15. **Vandenbroucke JP,** von Elm E, Altman DG, *et al*. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): explanation and elaboration. *PLoS Med* 2007;**4**:e297.