Carcinogenesis vol.28 no.8 pp.1774–1779, 2007
doi:10.1093/carcin/bgm129
Advance Access publication June 4, 2007

A bibliometric analysis of scientific production in cancer molecular epidemiology

Donatella Ugolini[1,2,*], Riccardo Puntoni[2], Frederica P.Perera[3], Paul A.Schulte[4] and Stefano Bonassi[5]

[1]Dipartimento di Oncologia, Biologia e Genetica, University of Genoa, Genoa 16132, Italy, [2]Units of Epidemiology and Biostatistics, National Cancer Research Institute, Genoa 16132, Italy, [3]Department of Environmental Health Sciences, Mailman School of Public Health of Columbia University, New York, NY 10032, USA, [4]Centers for Disease Control and Prevention, National Institute for Occupational Safety and Health, Cincinnati, OH 45226, USA and [5]Unit of Molecular Epidemiology, National Cancer Research Institute, Genoa 16132, Italy

[*]To whom correspondence should be addressed. Tel: +39 010 5600071; Fax: +39 010 5600501;
Email: donatella.ugolini@istge.it

***Objectives*: The main purpose of this research was to compare the scientific production in the field of cancer molecular epidemiology among countries and to evaluate the publication trend between 1995 and 2004. *Methods*: A bibliometric study was carried out searching the PubMed database with a combined search strategy based on the keywords listed in the medical subject headings and a free text search. Only articles from a representative subset of 92 journals—accounting for 80% of papers identified—were selected for the analysis, and the resulting 13 240 abstracts were manually checked according to a list of basic inclusion criteria. The study evaluated the number of publications and the impact factor (mean and sum), absolute and normalized by country population and gross domestic product. *Results*: A total of 3842 citations were finally selected for the analysis. Thirty-seven percent came from the European Union (UK, Germany, Italy, France and Sweden ranking at the top), 31.6% from USA and 9.7% from Japan. The highest mean impact factor was reported for Canada (6.3), USA (5.9), Finland (5.8) and UK (5.2). Finland, Sweden and Israel had the best ratio between scientific production and available resources. 'Genetic polymorphism, glutathione transferase, breast neoplasm, risk factors, case–control studies and polymerase chain reaction' were the most used keywords in each of the subgroups evaluated, although inclusion criteria may have privileged studies dealing with exogenous carcinogens. *Conclusion*: Cancer molecular epidemiology is an expanding area attracting an increasing interest. The identification of an operative definition is a necessary condition to give to this discipline a unique scientific identity.**

Introduction

Scientists have long recognized the intrinsic limitations of the traditional epidemiological design to discern the causal link between risk factors and disease occurrence in this evolving society. The pressing need of developing new tools for etiologic research was the driving force that in 1982 moved Perera and Weinstein to propose an enhancement of the epidemiological approach through 'the incorporation of laboratory analytical techniques to elucidate the biochemical or molecular basis of disease etiology' (1). Since then, many studies have been conducted to investigate the distribution of diseases in human populations and their determinants, incorporating molecular biology techniques into the epidemiologic design (2–4).

In the last decades, molecular epidemiology has gained a well-established position in the field of cancer research, with a number of dedicated researchers and institutions all over the world. This increased popularity has resulted in a growing scientific production, whose impact in the field is still to be fully quantified. Bibliometric studies are systematically conducted to evaluate the amount and the evolution of the scientific production among countries in major biomedical fields (5–12), but are particularly useful for novel disciplines, whose impact on the larger field of biomedical research has yet to be fully evaluated.

Bibliometry surveys the scientific production of a scientist, a research unit, an institution or a country by taking into consideration the historical development of a discipline or by quantifying its role in the domain of science, or prospectively, identifying research fronts. To perform this evaluation, citation analysis is currently used.

Citation analysis is defined as the number of times an article is cited as a reference in other articles and is based on the general assumption that the number of citations reflects an article's influence and notoriety and, hence, its quality. The databases most commonly used are those produced by the Thomson Scientific (formerly known as Thomson Institute for Scientific Information), which evaluates the papers published in >7500 peer-reviewed journals in the sciences and social sciences, and each year publishes an index (Journal Citation Reports) based on cited articles (13).

The main purpose of this paper is to provide a report on the scientific production in the field of cancer molecular epidemiology among countries. To this aim, the geographical distribution and the temporal trend of papers published between 1995 and 2004 have been investigated.

Basic information about published papers includes the list of those journals most often chosen by researchers in the field and further consideration was given to the impact factor (IF) of the journals where the papers were published. This parameter gives further information about the quality of the published material, especially if evaluated in the context of major socioeconomic variables, i.e. the source country population and its gross domestic product (GDP).

Finally, the evaluation of most frequently used keywords in cancer molecular epidemiology papers provided useful hints about the identification of main research trends and helped to interpret the perspective of evolution of this field.

Abbreviations: EU, European Union; IF, impact factor; MeSH, medical subject headings; mIF, mean impact factor.

Methods

Bibliographic search

The search for papers to be included in the analysis was performed using the PubMed database (National Library of Medicine, National Institutes of Health, Bethesda, MD—http://www.ncbi.nlm.nih.gov/sites/entrez).

The search strategy was built by (i) identifying, whenever possible, the keywords listed in the medical subject headings (MeSH) thesaurus [words appearing in the MeSH field (mesh)], i.e. the vocabulary of medical and scientific terms that are assigned to most PubMed documents by a team of trained experts (indexers) and (ii) performing, for search completeness, a free text search [words in title or abstract field (tiab)].

Our search covered the papers published during 1995–2004 and was performed on 30 June 2005. Because of the lack of specific keywords (MeSH terms) that could unequivocally identify cancer molecular epidemiology studies (14), the strategy adopted for this search was complex and included several keywords (MeSH terms) and also free text terms. The first group of keywords (MeSH terms) and free text terms refers to main concepts of molecular epidemiology such as 'epidemiology, molecular', 'biological markers', 'biomarkers', 'polymorphism, genetic', 'genotype', 'susceptibility', 'microarray'. This selection is extended through the 'OR' operator to specific biomarker name ('chromosome aberrations', 'micronucleus test', 'sister chromatid exchange', 'comet assay', etc.) as additional concepts, in order to broaden the search. The second group includes the concept of cancer ('neoplasms', 'carcinogens', 'mutagens', etc.) and of risk factors of environmental origin ('environmental pollution', 'occupational diseases', 'smoking', etc.). This latter sector is strictly related to cancer, but rarely indexed with a cancer-related keyword. The third are the keywords (MeSH terms) necessary to restrict the search to human studies ('human'). The last two groups of keywords (MeSH terms) are designed to exclude clinically oriented studies ('diagnosis', 'therapy', 'prognosis', 'survival', etc.) and non-research publications, such as reviews, letters, news, etc.

Keywords (MeSH terms) or free text terms related to epidemiological methods, e.g. case control, cohort, etc., have not been used since the result was too restrictive.

The study included all peer-reviewed papers with an abstract, and excluded reviews, news, congresses, case reports and letters to the editor [as identified in the publication type field (pt)].

The articles retrieved by our search strategy were manually reviewed before classifying them as a molecular epidemiology study of cancer. The basic inclusion criteria were the following: 'all studies focused on cancer whose methodology clearly present an epidemiologic study design and the use of a molecular biology methods'. Papers lacking a clear definition of the epidemiological study and of the laboratory technique were excluded, as were studies unrelated to cancer. Results of the *in vitro* challenge assays were also excluded. Studies on cancer patients (without controls) or molecular/cellular characterization of samples from cancer patients were included in the study only if an evaluation of risk factors through interview, questionnaire, etc. was present. Studies based on clinical output (diagnosis, therapy, prognosis, survival) were excluded in order to better limit the field. This latter exclusion criteria contributed to remove from the analysis many studies in fields greatly contributing to cancer molecular epidemiology such as those on infections agents.

Journals selection

Since our search strategy produced a very high number of articles, we selected the source journals with the purpose of producing a representative sample of the international scientific production in the field of cancer molecular epidemiology. Retrieved articles were published in >1000 journals, but only 330 of them published >3 articles during the period considered. Out of these, 97 journals that published 80% of the articles (each publishing >15 articles in these 10 years) were chosen. Five journals without IF were eliminated. The bibliometric analysis has been performed on the remaining 92 journals listed in Table I that were considered representative of cancer molecular epidemiology literature and sufficient to outline a trend.

Countries

The European Community [European Union (EU)] was defined as the 15 official member states plus Norway, given its inclusion in the European economic area and in all calculations concerning the EU issued by the Statistical Office of the European Communities. Papers from England, Scotland, Northern Ireland and Wales were grouped under the heading UK. For non-European countries, only data from 19 countries with >10 entries during 1995–2004 were evaluated.

The first author's country was considered as the country of origin of the article. Occasionally, it was necessary to manually identify the country source after consulting other bibliographic sources.

For each country, the number of publications and the mean impact factor (mIF) (sum of the IF divided by number of publications) were reported. To facilitate the comparability between countries, we eliminated the effects of the country size and the heterogeneous availability of resources. This was done by calculating the ratio between the scientific production of each country (expressed as the sum of the IF of all published papers) and the population size (number of inhabitants expressed in millions of inhabitants) or the national gross domestic product (expressed in current billion US dollars).

Demographic and economic data for each country were retrieved from Statistical Office of the European Communities or other international statistical reviews (15–17) and represent an average figure of the period under analysis.

Keywords

Keywords were defined as MeSH terms assigned to PubMed documents by a team of trained experts (indexers). In the indexing process, a variable number of terms (∼5–15) is assigned to each journal article to properly identify the content (18). The keywords (MeSH terms) used by PubMed experts to classify the 3842 articles selected for the study included as many as 3266 different terms. Of these, only 1792 (54.9%) were used more than twice. Keywords with similar meaning were assembled to produce a list of the most often used terms. Only keywords used >15 times, i.e. 725 (22.2%), were included in the analysis. The keywords were arbitrarily assembled in six groups identified by using higher order keywords in the MeSH tree structure used by indexers: genetic phenomena and processes (including 27% of all keywords), biomarkers (9%), neoplasms by sites (20%), environment and public health (20%), epidemiologic methods (16%) and laboratory techniques and procedures (8%).

Results

A total of 13 240 citations were retrieved from the PubMed database applying the search strategy reported above, and all corresponding abstracts were manually reviewed. A total of 3842 papers (29%) met the inclusion criteria reported in the methods session and were further evaluated.

Table I. List of the 92 journals selected for the bibliometric survey on cancer molecular epidemiology (Journals with IF publishing >15 papers on the topic)

Journals	% of articles
Cancer Epidemiol. Biomarkers Prev.	10.30
Cancer Res.	7.20
Mutat. Res.	6.70
Int. J. Cancer	6.40
Carcinogenesis	6.00
Br. J. Cancer	3.90
Cancer Lett.	3.70
J. Natl. Cancer Inst.	2.60
Clin. Cancer Res.	2.40
Pharmacogenetics	2.40
Cancer	2.20
Anticancer Res.	2.10
Cancer Genet. Cytogenet.	1.90
Am. J. Hum. Genet.	1.70
Oncogene	1.60
Environ. Mol. Mutagen.	1.50
Toxicol. Lett.	1.30
Blood	1.20
Br. J. Haematol.	1.10
J. Clin. Oncol.	1.10
Hum. Mutat.	1.00
Prostate	1.00

Other journals (publishing <1% of the papers selected) in alphabetical order: *American Journal of Epidemiology; American Journal of Medical Genetics; American Journal of Pathology; Annals of Human Genetics; Biochemical and Biophysical Research Communications; Breast Cancer Research; Breast Cancer Research and Treatment; Cancer Causes and Control; Cancer Detection and Prevention; Clinical Chemistry; Clinical Endocrinology (Oxf); Clinical Genetics; Cytogenetics and Cell Genetics; Diagnostic Molecular Pathology; European Journal of Cancer; European Journal of Cancer Prevention; European Journal of Endocrinology; European Journal of Human Genetics; Gastroenterology; Genes, Chromosomes and Cancer; Genetic Epidemiology; Genomics; Gut; Gynecologic Oncology; Haematologica; Hepatology; Human Genetics; Human and Molecular Genetics; Human Pathology; International Journal of Epidemiology; International Journal of Molecular Medicine; Internationl Journal of Oncology; International Journal of Radiation Oncology, Biology, Physics; JAMA, Journal of the America Medical Association; Japanese Journal of Cancer Research; Japanese Journal of Clinical Oncology; Journal of Biological Chemistry; Journal of Cancer Research and Clinical Oncology; Journal of Clinical Endocrinology and Metabolism; Journal of Clinical Pathology; Journal of Experimental and Clinical Cancer Research; Journal of Gastroenterology and Hepatology; Journal of Human Genetics; Journal of Immunology; Journal of Infectious Diseases; Journal of Investigative Dermatology; Journal of Medical Genetics; Journal of Medical Virology; Journal of Pathology; Journal of Urology; Laboratory Investigation; Lancet; Leukemia; Leukemia and Lymphoma; Leukemia Research; Lung Cancer; Medical and Pediatric Oncology; Melanoma Research; Modern Pathology; Molecular Carcinogenesis; Molecular and Cellular Probes; Nature Genetics; New England Journal of Medicine; Oncology; Oncology Reports; Oral Oncology; Proceedings of the National Academy of Sciences, USA; Science; Tissue Antigens; Urology*

Number of papers

The total number of published papers in the field of cancer molecular epidemiology during the 10 years period 1995–2004 increased from 369 in the biennium 1995–1996 to a maximum of 1233 papers in the biennium 2001–2002. In the last period there was a decrease with 916 published papers (Table II). To provide a comparison with the general trend of publication in cancer literature, the number of papers extracted with an automatic search tool built by PubMed experts and called 'cancer subset' (19) was reported in the same table.

All European countries except Luxembourg were represented. The most productive countries were the UK (20.8% of total European

Table II. Number of published papers in the field of cancer molecular epidemiology and mIF by country and year of publication (biennium)

Country	1995–1996		1997–1998		1999–2000		2001–2002		2003–2004		Total		
	Number	mIF	Number	mIF	Number	mIF	Number	mIF	Number	mIF	Number	mIF	mIF Rank
UK	37	4.5	36	4.5	70	5.4	97	6.1	55	5.3	295	5.2	2
Germany	25	4.6	25	2.5	53	4.4	80	5.0	46	4.4	229	4.2	7
Italy	22	2.7	31	3.6	44	4.6	52	4.7	47	4.8	196	4.1	8
France	8	3.6	16	3.1	30	5.2	37	5.0	42	5.1	133	4.4	5
Sweden	14	3.4	16	4.5	37	4.5	37	4.8	24	5.0	128	4.4	4
Spain	11	5.4	8	2.9	19	4.8	37	4.5	18	4.0	93	4.3	6
The Netherlands	10	3.9	11	4.7	18	4.8	27	6.8	24	5.3	90	5.1	3
Finland	8	7.5	6	4.6	14	5.5	22	5.7	14	5.6	64	5.8	1
Greece	3	1.3	3	2.7	8	2.8	13	2.8	14	3.1	41	2.5	14
Portugal	2	0.7	2	3.8	7	2.6	13	5.7	14	3.5	38	3.3	12
Denmark	6	2.5	4	4.0	8	3.8	9	4.2	10	4.8	37	3.8	9
Austria	1	0.8	1	1.5	2	6.7	11	3.4	14	5.6	29	3.6	11
Norway	8	2.2	6	3.7	5	5.8	4	4.4	1	2.3	24	3.7	10
Belgium	3	2.1	1	1.6	6	3.9	5	3.6	3	1.9	18	2.6	13
Ireland	2	1.3	0	0.0	1	1.5	3	5.6	0	0.0	6	1.7	15
Luxembourg	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	16
EUROPE	160	2.9	166	3.0	322	4.1	447	4.5	326	3.8	1421	3.7	
USA	113	6.1	150	6.3	264	5.9	391	6.2	298	5.3	1216	5.9	2
Japan	34	3.8	42	2.7	104	3.9	126	4.0	67	3.4	373	3.5	8
Taiwan	8	3.0	8	3.1	26	4.6	27	4.7	25	4.0	94	3.9	6
Australia	4	5.4	10	4.7	18	6.0	41	5.5	18	4.0	91	5.1	3
South Korea	3	2.0	3	2.2	14	3.9	25	3.7	35	3.7	80	3.1	11
China	2	1.0	9	3.3	10	4.1	28	4.9	30	5.2	79	3.7	7
Canada	3	3.7	4	8.4	19	6.5	31	7.3	20	5.6	77	6.3	1
India	11	2.2	9	2.3	13	5.9	12	2.1	14	3.4	59	3.2	10
Poland	2	4.9	5	2.1	11	3.1	9	2.9	16	4.3	43	3.5	9
Israel	3	3.6	3	3.2	10	5.8	14	6.9	6	5.4	36	5.0	5
Honk Kong	4	2.3	3	3.3	9	8.1	10	4.3	4	7.3	30	5.1	4
Turkey	4	3.9	8	1.6	7	4.6	5	2.1	6	1.7	30	2.8	13
Brazil	2	1.1	2	1.1	3	4.3	11	3.3	11	2.4	29	2.4	15
Czech Republic	2	1.7	6	1.8	3	3.2	8	3.6	5	2.3	24	2.5	14
Switzerland	0	0.0	4	4.9	6	5.4	12	3.7	0	0.0	22	2.8	12
Hungary	1	1.1	3	1.1	8	2.2	1	1.7	3	1.3	16	1.5	19
Russian Federation	2	1.8	2	1.1	4	2.5	3	2.6	3	1.4	14	1.9	17
New Zealand	1	1.5	2	1.2	3	3.1	3	1.8	1	2.3	10	2.0	16
Mexico	1	3.0	4	1.8	5	4.6	0	0.0	0	0.0	10	1.9	18
ALL	369	1.8	452	2.0	872	2.8	1233	3.0	916	2.5	3842	2.4	
Cancer literature	235 722		244 283		253 071		262 654		276 178		127 1908		

papers), Germany (16.1%) and Italy (13.8%) followed by France (9.4%), Sweden (9.0%) and Spain (6.5%). During the whole period, the EU published 1421 papers (37% of the total). In the same period, authors from USA produced 31.6% of the literature, Japan 9.7%, Taiwan and Australia 2.4% and South Korea and China 2.1%.

Quality of papers

The highest mean IFs were found for papers published by authors from Canada (6.3), USA (5.9), Australia (5.1), Hong Kong (5.1), Israel (5) and Taiwan (3.9). The overall performance of European papers was 3.7. Finland ranked first with a mean IF of 5.8 followed by the UK (5.2), The Netherlands (5.1), France and Sweden (4.4) and Spain (4.3). All other European countries had a mean IF between 4.2 and 1.7 (Table II). The global IF of all countries was 2.4.

Scientific production vis-à-vis population and gross domestic product

The ratio between the sum of IF and the resident population (expressed in millions of inhabitants), which describes the IF standardized by the size of the country, showed a mean value of 3.8 for Israel, 2.5 for USA and Australia, 2.4 for Hong Kong, 2.1 for Europe and 1.7 for Taiwan. In Europe, Finland ranked first (6.9), followed by Sweden (6.5), The Netherlands (3.1), the UK and Denmark (2.7) and Norway (2.0) (Figure 1).

The ratio between national IF and gross domestic product (expressed in current billion US dollars), which provides a resources corrected evaluation of the research quality, was particularly high for Israel (21.5), Taiwan (13.5), Czech Republic (12.2), Australia (12.0) and Hong Kong (9.9). The mean European score was 8.7, with the best performances in Finland (26.8), Sweden (23.1), The Netherlands and Portugal (11.6) and in the UK (10.9) (Figure 1).

Research topics

Table III gives the top 10 terms for each homogeneous groups of keywords (MeSH terms). In general, the most frequently used keywords were as follows: risk/risk factors (3192 times), genetic polymorphisms (2300), mutation (1793), genotype (1623), breast neoplasms (1506) and case–control studies (1432). Polymerase chain reaction (1113) was the most frequently used term among laboratory techniques, and glutathione transferase (911) among biomarkers.

Discussion

Our study shows that the two areas in the world with the highest scientific production in the field of cancer molecular epidemiology are Europe and USA. Taking into account the different size and availability of resources among countries, some areas of excellence emerge, such as Northern Europe and Israel. Among European countries, the analysis confirms the results observed in other biomedical disciplines, with the UK ranking first both in quantity and quality of scientific production. As a whole, the mean IF of cancer molecular

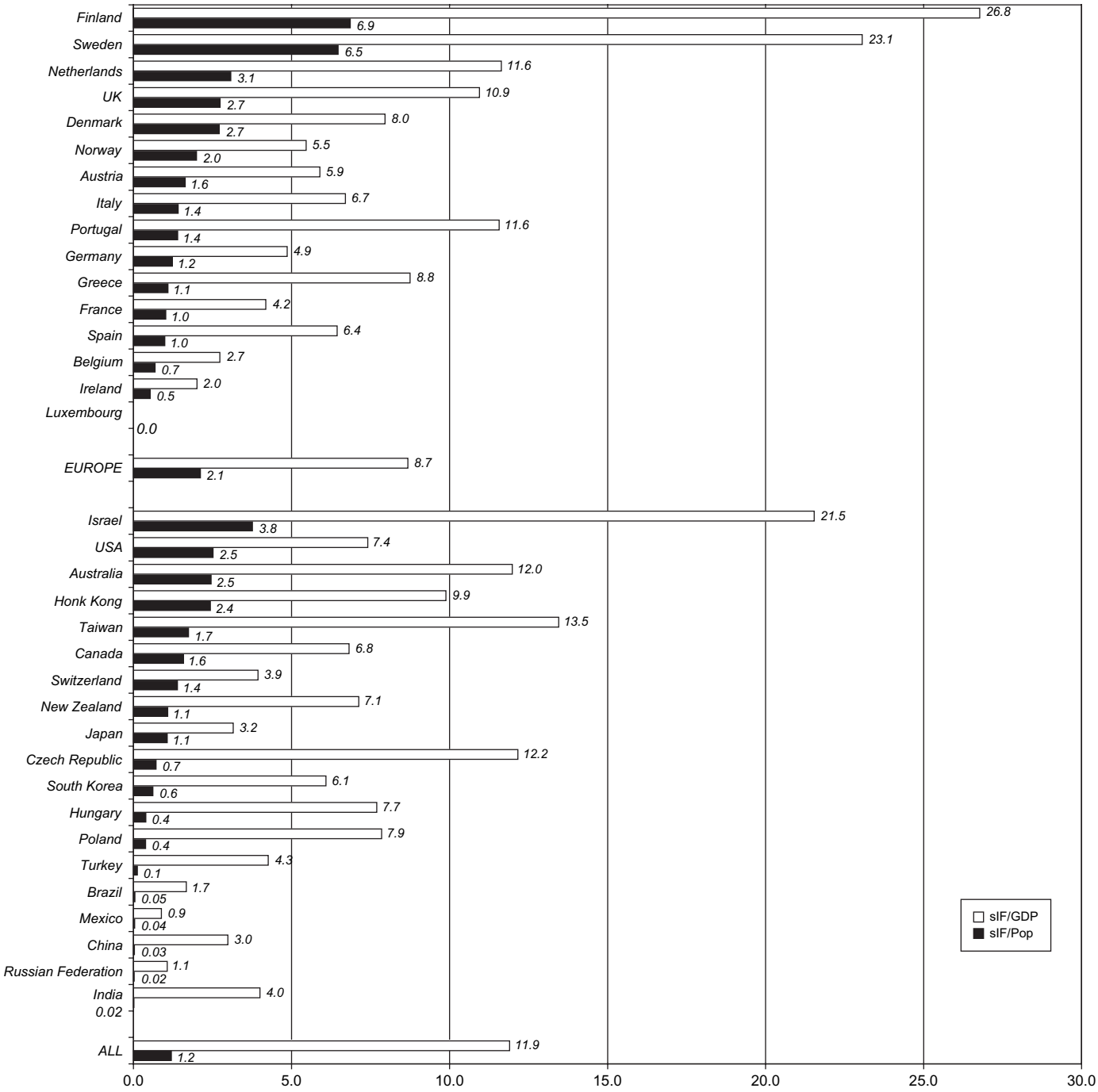

Fig. 1. Scientific production in the field of cancer molecular epidemiology in all countries standardized by population and economic parameters. Bars represent the ratio between scientific production of each country (sIF, expressed as the sum of the publications IF) and population (Pop, number of inhabitants expressed in millions of inhabitants) or economic parameters (gross domestic product (GDP), expressed in current billion US dollars).

epidemiology papers is higher than the corresponding value in oncology or other major fields of medicine, e.g. virology, rheumatology, ophthalmology, etc. (7–12).

During the first 8 years of the observation, the number of papers focused on cancer molecular epidemiology constantly increased all over the world. The decrease observed in the last biennium may be due (besides a possible decline of interest) to random variation in the process of abstracts manual selection (the total number of papers retrieved did not decrease accordingly) or to an incomplete collection of papers published in more recent years. However, despite this final discrepancy, most countries greatly increased their production during the years surveyed. Comparing the last biennium of the survey (2003–2004) to the first (1995–1996), the overall increase was 148%, which, if compared with the corresponding 17% of total cancer literature, provides a quantitative figure of the increasing interest for this field.

During the examined period, authors from all European countries published papers in cancer molecular epidemiology journals, with the exception of Luxembourg. Large countries, such as the UK, Germany, Italy and France published the highest number of papers. Among individual, non-EU countries (besides USA, that with the highest numbers of published papers (1216)), Japan (373), Taiwan (94) and Australia (91) were at the top.

Table III. List of the keywords (MeSH terms) most frequently assigned by PubMed indexers to papers in the field of cancer molecular epidemiology

Genetic phenomena and processes (27%)	Citations	**Biomarkers (9%)**	Citations	**Neoplasms by sites (20%)**	Citations
Polymorphism, Genetic	2300	Glutathione transferase	911	Breast neoplasms	1506
Mutation	1793	*p53* genes and protein	453	Lung neoplasms	1199
Genotype	1623	*Brca1/Brca2* genes and protein	386	Colorectal neoplasms	617
Genetic predisposition to disease	1046	Cytochrome P-450	374	Prostatic neoplasms	596
DNA damage and repair	605	Chromosome aberrations	204	Head and neck neoplasms	487
Heterozygote/homozygote	515	DNA adducts	170	Leukemia	339
Gene frequency	377	Arylamine *N*-acetyltransferase	143	Skin neoplasms and melanoma	337
Phenotype	321	Micronucleus tests	122	Liver neoplasms	315
Loss of heterozygosity	278	Sister chromatid exchange	103	Bladder neoplasms	291
Gene expression	211	*Ras* genes and protein	95	Ovarian neoplasms	244
Environment and public health (20%)	Citations	**Epidemiologic methods (16%)**	Citations	**Laboratory techniques and procedures (8%)**	Citations
Risk/risk factors	3192	Case-control studies	1432	Polymerase chain reaction	1113
Smoking	1107	Odds ratio	1003	DNA mutational analysis	341
Population	959	Comparative study	580	*In situ* hybridization, Fluorescence	218
Environmental exposure	403	Incidence	424	Pedigree	173
Occupational exposure/industry	417	Cohort studies	393	Sequence analysis	152
Air pollution	263	Prevalence	325	Amino acid substitution	124
Drinking/alcohol	203	Multivariate analysis	274	Electrophoresis	117
Diet	191	Confidence intervals	192	Chromosome mapping	102
Risk assessment	188	Regression analysis	189	Immunoblotting	92
Public health	179	Reference values	148	Flow cytometry	39

Nation rankings changed considerably when other end points were considered, such as the mean IF or the sum of IF adjusted by number of inhabitants or by GDP. The results of this survey extend to the field of cancer molecular epidemiology the common finding that small countries usually perform better than larger ones when the quality of the scientific production is considered (20). A better utilization of resources and a higher proportion of the GDP assigned to research are the most likely explanations (21–23).

The interpretation of the results of this bibliometric study should take into account a number of potential limitations. The most remarkable that may have affected the search strategy is the lack of standardized concepts to quantify the scientific production in the field of cancer molecular epidemiology. For example, it should be highlighted that some molecular epidemiology studies on biomarkers of infection, immunology, hormones, inflammation, etc. were not included in the analysis due to the lack of specific keywords in these fields. Furthermore, these biomarkers were commonly used in clinical studies, which were excluded from our study according to the search criteria. This potential bias, which is unlikely to occur differentially by country and calendar year, was addressed using an extensive choice of keywords and free text terms.

Another potential source of bias is the manual selection performed by three of the authors (D.U., R.P. and S.B.) of all candidate papers retrieved from the PubMed database (13 240!); despite the previous standardization among evaluators and the extensive quality controls, this procedure may have generated some discrepancy. Finally, the PubMed database is biased in favor of English-language journals; thus our survey may have penalized those countries that have a tradition of publishing in their own language journals. It is possible that some countries more than others, e.g. Japan and Russia, suffered particularly in this respect.

Problems were also encountered in the identification of main author address, i.e. the institutional and geographical affiliation. If the author's address is reported inaccurately, a margin of error in data extraction is possible. Furthermore, this approach does not adequately reflect the contribution of various countries to international collaborative studies (e.g. large pooled analyses); however, the present analysis is based on large numbers, and international collaborative studies often entail a rotation of the first author.

An additional limitation of the study—which affects all bibliometric studies—is represented by the intrinsic inaccuracy in the measure used to describe the quality of scientific production. The IF of a journal is the average number of citations that a paper published in that journal receives in the 2 years following publication. Clearly, this index does not give a score of the single paper, but is a journal average value, and may be severely conditioned by the ups and downs of scientific interests. This issue is currently the focus of a debate within research evaluators and funding agencies about the best methods for the allocation of resources (24–26). The use of the citation frequency to measure the impact of a published paper is the most accessible and suitable source of data for the evaluation of the scientific production. However, this approach is not flawless either, and ideally, an exhaustive survey would combine data of different bibliometric indicators.

A descriptive analysis comparing nations is an essential step in understanding science policies and a source of beneficial information that enables a country to define its position with respect to competitors. This in turn allows for better exploitation of opportunities arising in all scientific fields. These surveys offer a broad review of the existing data and help to gather impressions of scientific publication trends and the visibility of a country's production.

A further consideration regards a problem that affects many biomedical disciplines. The analysis of keywords revealed a high heterogeneity of terms. In fact, only 22.2% of keywords are cited >15 times and 54.9% more than twice. Our analysis re-elaborated all keywords in order to group together concepts expressed with similar terms to offer a more synthetic picture of current research trends within cancer molecular epidemiology.

This report represents the first effort to explore the geographical distribution and the development in the field of cancer molecular epidemiology. This exercise is especially useful for young disciplines since it provides quantitative information about the growth of the field, the geographical distribution of the scientific excellence and through the use of keywords, a ranking of the most successful topics. Concerning cancer molecular epidemiology, there is an additional need of characterizing the whole discipline, which nowadays is lacking not only a formal definition but also an operative definition that can unequivocally identify the boundaries of the discipline.

Funding

Associazione Italiana per la Ricerca sul Cancro; Agenzia Spaziale Italiana, Italy.

Acknowledgements

We thank Roberto Garrucciu, Genoa, Italy, for his support in data collection and computer analysis and Barbara Landreth, Cincinnati, OH, for her critical review of the paper.

Conflicts of interest: None declared.

References

1. Perera,F.P. *et al*. (1982) Molecular epidemiology and carcinogen-DNA adduct detection: new approaches to studies of human cancer causation. *J. Chronic Dis.*, **35**, 581–600.
2. Sculte,P.A. and Perera,F.P. (eds.) 1993 *Molecular Epidemiology: Principles and Practices*. Academic Press, New York, NY. 1993.
3. Perera,F.P. *et al*. (2000) Molecular epidemiology: recent advances and future directions. *Carcinogenesis*, **21**, 517–524.
4. Bonassi,S. *et al*. (2005) Human population studies with cytogenetic biomarkers: review of the literature and future prospectives. *Environ. Mol. Mutagen.*, **45**, 258–270.
5. Michalopoulos,A. *et al*. (2005) A bibliometric analysis of global research production in respiratory medicine. *Chest*, **128**, 3993–3998.
6. Rahman,M. *et al*. (2005) Research articles published in clinical radiology journals: trend of contribution from different countries. *Acad. Radiol.*, **12**, 825–829.
7. Falagas,M.E. *et al*. (2005) Estimates of global research productivity in virology. *J. Med. Virol.*, **76**, 229–233.
8. Cimmino,M.A. *et al*. (2005) Trends in otolaryngology research during the period 1995–2000: a bibliometric approach. *Otolaryngol. Head Neck Surg.*, **132**, 295–302.
9. Ugolini,D. *et al*. (2003) Oncological research overview in the European Union. A 5-year survey. *Eur. J. Cancer*, **39**, 1888–1894.
10. Grossi,F. *et al*. (2003) Geography of clinical cancer research publications from 1995 to 1999. *Eur. J. Cancer*, **39**, 106–111.
11. Ugolini,D. *et al*. (2001) How the European Union writes about ophthalmology. *Scientometrics*, **52**, 45–58.
12. Mela,G.S. *et al*. (1998) An overview of rheumatological research in the European Union. *Ann. Rheum. Dis.*, **57**, 643–647.
13. Institute for Scientific Information (1997) *SCI: Science Citation Index-Journal Citation Reports*. The Institute for Scientific Information, Philadelphia, PA.
14. Ugolini,D. *et al*. (2006) Searching PubMed for molecular epidemiology studies: the case of chromosome aberrations. *Environ. Mol. Mutagen.*, **47**, 227–229.
15. Eurostat. *Research and Development: Annual Statistics*. Statistical Yearbook. Statistical Office of the European Communities Luxembourg, Belgium. 1996–2001.
16. International Monetary Fund, 700 19th St. NW, Washington, DC 20431. Available from: http://www.imf.org. Last accessed date: October 2005.
17. The United Nations Department of Economic and Social Affairs, Population Division, United Nations Statistics Division, New York, NY 10017 United States of America. Available from: http://unstats.un.org/unsd/default.htm. Last accessed date: October 2005.
18. United States, National Library of Medicine, National Institute of Health. PubMed manual. NLM,8600 Rockville Pike, Bethesda, MD 20894. 2005. Available from: http://www.nlm.nih.gov/pubs/manuals/pm_workbook.pdf. Last accessed date: March 2006.
19. United States, National Library of Medicine, National Institute of Health. Search strategy used to create the cancer subset on PubMed. Strategy last modified February 2006. Available from: http://www.nlm.nih.gov/bsd/pubmed_subsets/cancer_strategy.html. Last accessed date: March 2006.
20. King,D.A. (2004) The scientific impact of nations. *Nature*, **430**, 311–316.
21. Rahman,M. *et al*. (2003) Biomedical research productivity: factors across the countries. *Int. J. Technol. Assess Health Care*, **19**, 249–252.
22. Benzer,A. *et al*. (1993) Geographical analysis of medical publications in 1990. *Lancet*, **341**, 247.
23. Anderson,A. (1992) Science in Europe. *Science*, **256**: 472.
24. Garfield,E. (2006) The history and meaning of the journal impact factor. *JAMA*, **295**, 90–93.
25. Hakansson,A. (2005) The impact factor—a dubious measure of scientific quality. *Scand. J. Prim. Health Care*, **23**, 193–194.
26. Kam,P.C. (2005) Impact factor: overrated and misused? *Anaesth. Intensive Care*, **33**, 565–566.

Received March 13, 2007; revised May 15, 2007; accepted May 20, 2007