Information measurement methods in applied research in subject services—Based on PubMed database data from 2018 to 2023

Jing Wang[*]

Library, Zhaoqing University, Zhaoqing, China
[*]Corresponding author

Abstract: *In this article, the author uses bibliometric theory and R software to take scientific research literature data in the field of empirical likelihood research in the PubMed database from 2018 to 2023 as the research object. Perspective conducted knowledge measurement analysis and visual analysis to sort out the evolution of research topics in this field in the past five years and provide a reference for researchers in related disciplines.*

Keywords: *Information measurement; PubMed database; Subject service; Visual analysis*

1. Introduction

The quantitative research method of scientific knowledge is a scientific research method created by Price who is the founder of American scientometrics and information science, in the 1920s. It is used to study the development of science itself. The scientific research paper retrieval system (Science Citation Index), which is well known by research scholars, has become an important tool for literature retrieval and citation analysis. The scientific research paper retrieval system has made fundamental contributions to the development of bibliometrics and scientometrics [1-3]. Scientometrics has developed rapidly with the development of subject librarian services. Subject librarian services have rapidly developed from traditional library and information services to various fields of scientific research. Subject librarian services have been deeply integrated with information science, resulting in the emergence of new interdisciplinary information metrology. At present, informetrics has been widely used in scientific evaluation indicators, database construction and information management systems. With the development of informetrics, many domestic scholars have investigated the development trends of scientometrics and informetrics from different angles [4-5]; the research hot spots or development trends of informetrics can be found through relevant literature [6-10]; China's Scientometrics Index Construction [11-15], etc. In order to help scientific researchers understand the current research status of empirical likelihood methods in the field of statistics and provide a reference for accurately establishing research directions, this article uses bibliometric methods to conduct a scientometric analysis of the literature data related to empirical likelihood research downloaded from the PubMed database.

2. Preparation of data set

In this article, the internationally mainstream authoritative database PubMed for biomedical research is selected as the data source. The author applies the professional search tool provided by the PubMed database to use "empirical likelihood" as the keyword, and the search time is January 1, 2018 - April 14, 2023. A total of 6434 records were retrieved. The R software is used to perform statistical analysis, visual analysis and econometric analysis on the retrieved 6434 literature data. The article mainly performs a quantitative analysis on the literature data from the aspects of additional keywords (ID), author keywords (DE), abstract (AB), number of articles published by the author, and country and institution, etc., in order to grasp the current status of censored data research in survival analysis. Trends, hot spots and development trends provide to some relevant researchers with research guidelines in this fiel. Then the results will improve the scientificity and foresight of researchers' research topics.

3. Descriptive statistical analysis

3.1 Basic descriptive statistical analysis results

The biblioAnalysis function of R software was used to perform descriptive analysis on the data downloaded from the PubMed database. From the analysis results, it can be seen that there are a total of 11,553 additional keywords (ID) in the literature data in the past 5 years, as well as the author keywords (DE). There are 11,553 and there are 28,908 authors. Authors appeared 35,487 times in all documents, including 315 authors of sole articles and 6,119 articles with multiple authors. In terms of author collaboration, there were 332 single-author articles. Each author published an average of 0.223 articles in the past five years. The average number of authors per article was 5.51; the number of co-authors per article in the five years was 4.49; the total number of keywords is 11,553. It can be seen that keywords are very widespread among empirical likelihood researchers. The average number of keywords per article is 1.79, which shows that research in this field is still in a relatively active research cycle in the past five years. These situations prompt researchers to determine their own research directions and hot spots based on currently active keywords. The statistical results show that 28,908 authors have published articles. It shows that this field is still a research hotspot in statistics. Many scholars are invested in this research field, the degree and scope of cooperation between researchers is very extensive, because Only 315 articles were by a single author. These situations also show that with the development of big data, the hot spots and depth of statistical research are becoming more and more in-depth, and most topics cannot be completed by a single author. Cooperation among researchers must be strengthened, especially the integration and development of interdisciplinary research.

3.2 Distribution of authors' countries

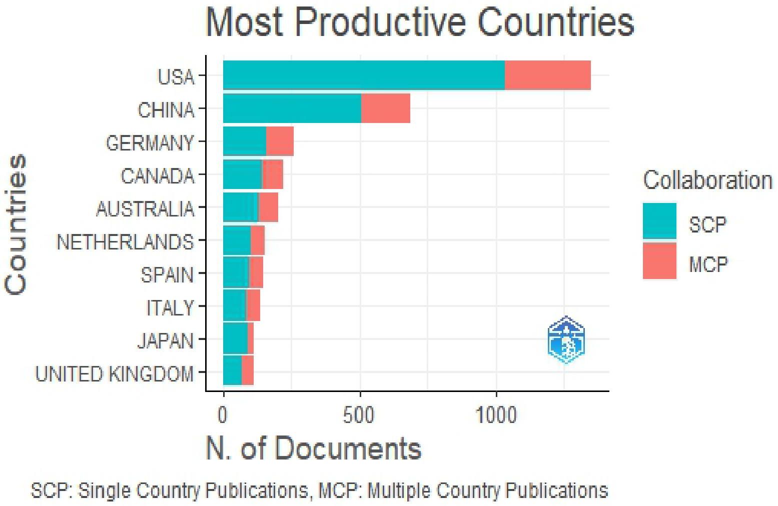

Figure 1: Corresponding author country distribution chart (Most Productive Countries)

From the statistical data Table 1 and statistical chart Figure 1 of the number of articles published by the corresponding author, it can be seen that the country with the highest number of articles published by the corresponding author is the United States, with 1,348 articles. Among them, there are 1,037 articles whose authors are all American authors, and 3,116 articles with non-American authors. The International cooperation rate is 0.231; China ranks second with 687 articles, including 504 articles whose authors are all Chinese authors, and 183 articles with non-Chinese authors. The international cooperation rate of Chinese authors is 0266; the total number of published articles ranks is third. Germany ranked first with 25 articles, of which 160 articles were all authored by German authors, and 96 articles were collaborated with non-German authors. The international cooperation rate of German authors was 0.375. It can be seen that the United States is still the center of empirical likelihood non-parametric statistical research. At the same time, China, Germany, Canada, and Australia have gradually become hot research countries in this field in recent years, and are forced that cannot be ignored in this field. The total number of articles published by the top five countries accounts is nearly 50% of the world's total. It can be seen that the total number of publications by Chinese authors has reached half of that of American authors, and the international cooperation rate of Chinese authors is higher than that of the United States.

Table 1: Corresponding Author's Countries

Serial Number	Country	Articles	SCP	MCP	MCP_ Ratio	Serial Number	Country	Articles	SCP	MCP	MCP_ Ratio
1	USA	1348	1037	311	0.231	21	DENMARK	36	20	16	0.444
2	CHINA	687	504	183	0.266	22	TURKEY	35	23	12	0.343
3	GERMANY	256	160	96	0.375	23	SOUTH AFRICA	32	15	17	0.531
4	CANADA	222	141	81	0.365	24	IRELAND	31	13	18	0.581
5	AUSTRALIA	203	130	73	0.36	25	PORTUGAL	29	18	11	0.379
6	NETHERLANDS	154	101	53	0.344	26	PAKISTAN	28	12	16	0.571
7	SPAIN	144	98	46	0.319	27	THAILAND	28	22	6	0.214
8	ITALY	133	86	47	0.353	28	SINGAPORE	26	17	9	0.346
9	JAPAN	112	88	24	0.214	29	NEW ZEALAND	24	12	12	0.5
10	UNITED KINGDOM	111	67	44	0.396	30	GREECE	22	11	11	0.5
11	FRANCE	108	75	33	0.306	31	POLAND	21	16	5	0.238
12	SWITZERLAND	105	35	70	0.667	32	MALAYSIA	18	11	7	0.389
13	KOREA	87	69	18	0.207	33	AUSTRIA	17	11	6	0.353
14	BRAZIL	79	63	16	0.203	34	SAUDI ARABIA	17	8	9	0.529
15	INDIA	65	53	12	0.185	35	FINLAND	15	5	10	0.667
16	SWEDEN	63	43	20	0.317	36	ARGENTINA	14	8	6	0.429
17	ISRAEL	61	45	16	0.262	37	CHILE	14	10	4	0.286
18	IRAN	52	36	16	0.308	38	COLOMBIA	14	12	2	0.143
19	NORWAY	47	21	26	0.553	39	GHANA	14	10	4	0.286
20	BELGIUM	42	16	26	0.619	40	NIGERIA	14	9	5	0.357

SCP: number of co-authored papers with authors of the same nationality, MCP: number of co-authored papers with authors from other countries, MCP_Ratio: ratio of international collaboration

3.3 Keyword analysis of published articles

Table 2: List of keywords for published articles (Most Relevant Keywords)

Serial Number	Author Keywords(ID)	Articles	Serial Number	Author Keywords(ID)	Articles
1	HUMANS	4521	21	ANTI-BACTERIAL AGENTS/THERAPEUTIC USE	286
2	FEMALE	1978	22	CROSS-SECTIONAL STUDIES	279
3	MALE	1786	23	PROBABILITY	272
4	RISK FACTORS	1425	24	SURVEYS AND QUESTIONNAIRES	261
5	ADULT	1186	25	PROSPECTIVE STUDIES	243
6	BAYES THEOREM	1100	26	CHILD PRESCHOOL	215
7	MIDDLE AGED	962	27	INFANT	209
8	AGED	798	28	TREATMENT OUTCOME	207
9	ADOLESCENT	637	29	LIKELIHOOD FUNCTIONS	195
10	YOUNG ADULT	587	30	UNITED STATES	195
11	RETROSPECTIVE STUDIES	525	31	INFANT NEWBORN	194
12	CHILD	448	32	COHORT STUDIES	189
13	ANIMALS	446	33	PREVALENCE	177
14	RISK ASSESSMENT	404	34	CHINA	172
15	COMPUTER SIMULATION	396	35	MODELS THEORETICAL	157
16	ALGORITHMS	365	36	INCIDENCE	156
17	AGED 80 AND OVER	325	37	TIME FACTORS	156
18	UNCERTAINTY	313	38	REPRODUCIBILITY OF RESULTS	150
19	MODELS STATISTICAL	306	39	PHYLOGENY	140
20	LOGISTIC MODELS	291	40	PREGNANCY	137

From Table 2, it can be seen from the top 40 keyword statistics of published data that the empirical likelihood non-parametric research method has been applied in various fields of statistics, such as: Anthropology (HUMANS), Women's Studies (FEMALE), Risk Factors (RISK FACTORS), Bayesian Theory (BAYES THEOREM), etc. Analyzing the specific situation, according to the statistics of published articles in the past five years, it can be seen that the author's keywords appear most often in anthropology, with a total of 4,521 documents using this keyword as a keyword; in second place is women's studies, with a total of 1,978 documents which is used as a keyword; in the third place is andrology, with a total of 1786 documents using this as a keyword; in the fourth place is risk factors, with a total of 1425 documents using this as a keyword; the fifth place is adult studies, and there are a total of 1186 documents with this keyword. Anthropology, which ranks first among a total of 6,434 documents, has 4,521 documents with this keyword as a keyword. It can be seen that anthropology is the most important research topic in empirical likelihood method research. The 10th-ranked keyword (YOUNG ADULT) appears 587 times, with an average of 11 documents using this keyword as the keyword. The statistical results show that keywords with the top 10 frequencies are the main direction in the field of empirical likelihood method research and deserve the attention of relevant researchers. Researchers can use this quantitative analysis information to find their own research topics more easily, avoid unnecessary document tracking processes, and save valuable scientific research time and funds.

3.4 Lokta analysis

Lokta analysis is a statistical analysis method given by Lotka who is a statistician for the American insurance company, in 1926. This method analyzes the production capacity of scientific, technological workers, their contribution to scientific, technological progress and social development through the statistics of published papers. Through statistical analysis with R software, it can be seen that the Lokta parameters corresponding to this research topic are $\beta = 2.463, C = 0.192$ from Figure 2. The significance P value of the test is 0.0001. In order to facilitate the intuitive understanding of Lokta analysis, from the "Scientific Productivity" distribution fitting in Figure 3, it can be seen that the fitting effect between the theory and the actual value of the Lokta distribution is quite good in this study. It is indicating that the scientific and technological production capacity transformation of relevant researchers is related. It's relatively obvious. From the statistical test, the significance level P value of 0.0001 is far smaller than the significance level. From a statistical point of view, the theoretical value and the actual value have a good fitting effect and they are statistically significant.

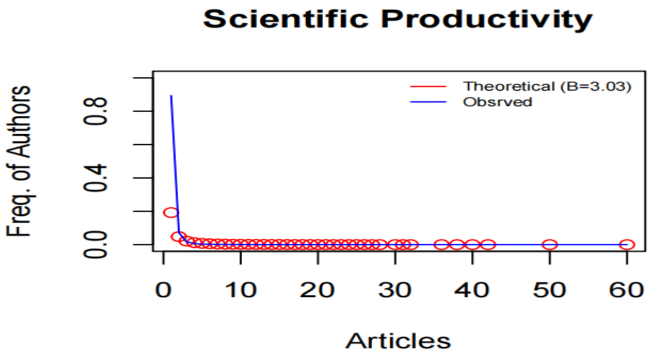

Figure 2: Scientific Productivity Distribution Fitted Plot

3.5 Visual analysis of author keyword cooperation network

Cooperation network analysis is a common method to study the activity of a specific research field. This method can provide research hot spots in related fields through diagrams and author collaboration network analysis, and then point out prominent research centers in related fields. The significance of the network is that the larger the network nodes, the more collaborators in the research field, indicating the higher the research activity in the field; the more connections, the more frequent collaborative research in the research fields at both ends of the connection. R software is used to visually analyze the author's keyword collaboration network. Through the collaborator network visualization diagram in Figure 3, it can be seen that the top-ranked keywordsalso have a relatively obvious collaboration network in terms of research cooperation in the past five years. Among the top 10 keywords, there are at least 10 or more research cooperation network connections, especially the keywords: HUMANS, FEMALE, MALE, RISK FACTORS, and ADULT. There are no less than 30 cooperation networks indicating research in these related fields. The cooperative research is very active and basically forms their own research centers. The remaining keyword cooperation networks are also very dense and have basically formed their own research centers. The cooperation network diagram shows that in scientific research with keywords as the hub, scientific research has increasingly become cross-scientific research, and early single research has completely failed to meet the needs of scientific development.

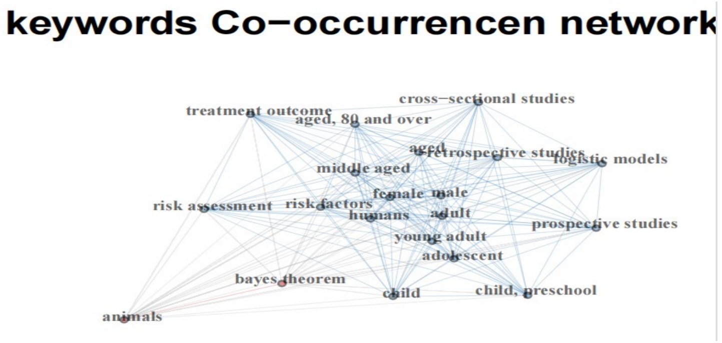

Figure 3: Collaboration Network Visualization

3.6 Semantic map analysis

Strategic coordinate map analysis in semantic map analysis is a two-digit graph established by centripetalness (X) and density (Y), which is used to measure the internal development of each topic in a certain research field and the mutual influence between topics. The density (Y) is used to evaluate the strength of the internal correlation of the topic, indicating the maintenance and development ability of the topic class itself; centripetal degree (X) is used to evaluate the closeness between one topic class and another topic class, the greater the centripetal degree. The closer the connection between topics, the more central the topic is in a certain research field. It can be seen from the strategic coordinate map in Figure 4 that the keywords BAYES THEOREM, HUMAN and ANIMALS each forms a research topic category in terms of centripetal degree, and each becomes the core of this research field. This strategic map can clearly see the development trends of theme categories.

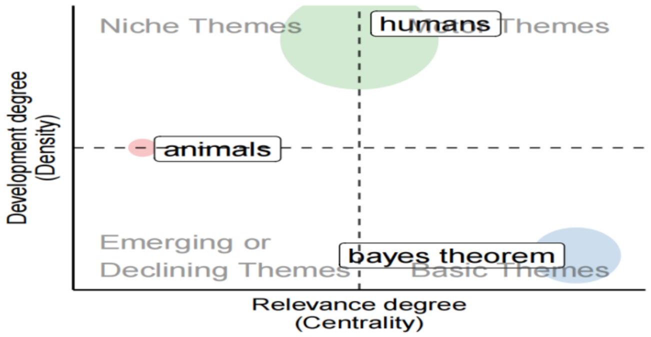

Figure 4: Strategic Coordinate Map

3.7 Keyword cloud analysis

From the keyword word cloud Figure 5, we can see that the keyword in the first topic is HUMANS, and keywords such as MALE and RISK FACTORS are ranked second and third respectively. The word cloud diagram more vividly presents the main hot spots and development directions in this research field to researchers. The result of BAYES THEOREM, which is in fourth place, is obviously not unexpected. The BAYES method has been proposed for nearly a hundred years. However, due to the subjectivity of prior information and the complexity of posterior distribution calculations, this method has great constraints and restrictions on specific use. It was rarely studied and adopted by statisticians at the beginning, but with the rapid development of personal computer software and big data, the two shortcomings of the BAYES method that were previously considered to be its own advantages and have become its own advantages. Researchers can make good use of proactive research and a posteriori update of the understanding of the problem, to find solutions to problems quickly. Therefore, the BAYES THEOREM method has now become an indispensable basic theoretical support for big data, especially machine learning.

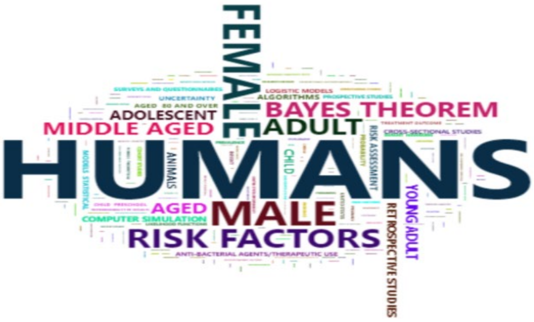

Figure 5: Keyword word cloud map

4. Conclusion

Big data has become the new "oil" of social development. With the emergence of massive data, statistics faces many challenges and development opportunities such as the reconstruction of its own theoretical framework and the universality of methods. In a chaotic research ecosystem, it is particularly important for researchers to find their own research footing. This work will help scholars quickly enter the research lane, produce research results that meet social needs, and promote the development of related science and technology. The scientometric analysis of knowledge given in this article can provide a reference for scholars. Through analysis, the paper provides research hotspots in the field of non-parametric empirical likelihood and sorts out the development trends in this field in the past five years. Knowledge graph combing and analysis provides more effective help for scientific researchers to effectively establish their own research fields and conduct efficient literature tracking.

Acknowledgement

This paper was supported by the Guangdong Province General University Characteristic Innovation Project (2022KTSCX150), Zhaoqing City Social Science Planning Project (23GJ-43), Zhaoqing Science and Technology Innovation Guidance Project (2023040317006), Key Project of Zhaoqing Institute of Education (ZQJYY2023021) and Zhaoqing College Quality Project and Teaching Reform Project (zlgc202112).

References

[1] Chu Jiewang, Sun Xiaoning. Analysis of the literature quantity of the current research status of domestic library knowledge management [J]. Library Theory and Practice, 2012(9): 21-26. (In Chinese)

[2] Hou Haiyan, Liu Zeyuan, Luan Chunjuan. Quantitative analysis of research frontiers of the International Institute of Scientometrics based on knowledge graph [J]. Scientific Research Management, 2009, 30(1): 164-170. (In Chinese)

[3] Jin Bihui, Zhang Jianggong, Chen Dingquan, et al. Development of China's scientific measurement indicators (CSI) [J]. Scientometrics, 2002, 54(1): 145-154. (In Chinese)

[4] Liang Liming. Scientometrics and Informetrics: Looking at China from the World [J]. Scientific Research Management, 2000, 21(3) 95-101. (In Chinese)

[5] Liang Guoqiang. Review of domestic bibliometrics [J]. Scientific and Technological Document Information Management, 2013(4): 58-59, 62. (In Chinese)

[6] Li Fengzhi. The definition and role of citation analysis [J]. Science and Technology Information, 2015(10):25-253. (In Chinaese)

[7] Li Weichao, Guo Jun, Li Jingyu. Comparison of domestic and foreign library and information science research hotspots - based on iConference [J]. New Century Library, 2021(3): 12-17. (In Chinaese)

[8] Lin Lili, Ma Xiufeng. Discovery and evolution analysis of domestic library and information science research topics based on LDA model [J]. Information Science, 2019()12: 87-92. (In Chinese)

[9] Hou Jianhua, Yang Xiucai, Zhou Lijuan. Analysis of cutting-edge themes and evolutionary trends in international library and information research [J]. Library and Information Service, 2016(13): 82-90. (In Chinese)

[10] Ma Feng, Yang Siluo. Comparative analysis of the theme content of domestic and foreign library and information science research [J]. Information Science, 2015(9): 140-145. (In Chinese)

[11] Teng Guangqing, Mou Dongmei, Ren Jing. Research on the application of foreign social network analysis in the field of bibliometrics [J]. Information and Documentation Work, 2014(1): 47-51. (In Chinese)

[12] Si Li, et al. Comparative analysis of research hot spots in library, information and archives management at home and abroad from 2014 to 2018 [J]. Library and Information, 2020(1): 75-82. (In Chinese)

[13] Su Xinning, Xia Lixin. Analysis of the subject areas of digital library research in my country from 2000 to 2009—based on CSSCI keyword statistics [J]. Journal of Library Science in China, 2011, 37(7):60-69. (In Chinese)

[14] Shi Yanqing, Sun Jianjun. Analysis of the research theme characteristics of my country's library and information disciplines in the world [J]. Library and Information Service, 2018(7): 66-76. (In Chinese)

[15] Wang Wei, Wang Liwei, Zhu Hong. Analysis of international informetrics research frontiers and hot spots [J]. Journal of Medical Informatics, 2010, 31 (2): 1-4. (In Chinese)