ISSN 2616-5775 Vol. 7, Issue 1: 28-34, DOI: 10.25236/AJCIS.2024.070105

Cross-application of Information Measurement Methods in the Field of Library Intelligence to Disciplinary Services-Based on PubMed Database Data from 2017-2022

Jing Wang*

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

Abstract: *In this paper, the author uses bibliometric theory and the open source software R software for data analysis to study the research literature data in the field of missing data research in PubMed database from 2017-2022. The author conducts the knowledge measurement analysis as well as visualization analysis from several perspectives, such as the number of articles, citations, collaborations, core authors, and research hotspots, in order to sort out the research themes in this field have evolved in the past five years, and provide reference for researchers in related disciplines.*

Keywords: *library intelligence, econometric analysis, PubMed database, knowledge measurement analysis, visualization analysis*

1. Introduction

The founders of American scientometrics and one of the founders of intelligence science, price and others, pioneered quantitative research methods to study science itself. The SCI research paper search system developed by Garfield, the founder of the Science Citation Index, became an important tool for literature search and citation analysis, and made an important contribution to the development of bibliometrics and scientometrics [4-5]. In recent years, scientometrics has developed rapidly and expanded from the traditional library and intelligence field to many fields of scientific research, and formed an interdisciplinary informatics with information science. At present, informetrics has been applied to database construction, information management systems, scientific evaluation indicators and scientific research information services. With the development of national informetrics, many domestic scholars have examined the development trend of scientometrics and informetrics from different perspectives [1-3], including the research hotspots or development trend of informetrics [6-11], the construction of Chinese scientometrics indexes [12-16], etc. With the continuous expansion of scientific research fields and the arrival of big data, scientific researchers need to regularly track research hotspots and sort out development trends in related fields, which brings a lot of troubles to researchers due to the ponderous data. At the same time, it also brings opportunities to the personnel related to the subject services of university libraries, which makes the subject librarian services find the areas where they can play their strengths, such as cooperating with professional researchers to conduct scientometric analysis in specialized fields. The author provides a comprehensive review of research in the field of missing data in order to provide valuable references for relevant scientific researchers to accurately establish research directions. In this paper, the author adopts bibliometric methods to conduct scientometric analysis of the relevant literature data downloaded from PubMed database.

2. Preparation of data set

In this paper, PubMed, the international mainstream authoritative database of biomedical research, was selected as the data source, and the professional search tool provided by PubMed database was used with "censored data" as the keyword, and the search time was from January 1, 2017 to May 1, 2022, and a total of 3733 records were retrieved. Among them, there were books and documents that did not meet the requirements of the subsequent analysis, and they were deleted, and the final valid data were 2810 records. Statistical analysis, visualization analysis, and econometric analysis of the retrieved 2810 literature data were performed using the open source software for data analysis, R

ISSN 2616-5775 Vol. 7, Issue 1: 28-34, DOI: 10.25236/AJCIS.2024.070105

software. The article mainly analyzed the literature data in terms of additional keywords (ID), author keywords (DE), abstract (AB), author's publication volume, and country and institution, in order to grasp the current trends, hot spots and development trends of the research on deletion data in survival analysis, to provide relevant researchers with research guidelines in this field, and to improve the scientific and forward-looking of researchers' research topics.

3. Descriptive statistical analysis

3.1 Results of basic descriptive statistical analysis

The bibliometric analysis function of R software was used for descriptive analysis of the data downloaded from the PubMed database. From the analysis results, it can be seen that there are a total of 6183 additional keywords (ID), and the same 6183 author keywords (DE) with 14603 authors in the literature data of the last 5 years. Authors appear 18,694 times in all the literature, including 63 authors of single-authored articles and 14,540 multi-authored articles. The total number of keywords is 6183, which shows that the keywords in the censored data are very extensive, and the average number of keywords per article is 2.5. The average number of keywords per article is 2.5, which indicates that the research in this field has been in a relatively active research cycle in the past five years. This suggests that researchers can determine their research directions and hotspots based on the current active keywords. The fact that 14,603 authors have published articles indicates that this field is still a hot spot in statistics, with a large number of scholars devoted to this area of research, and the extent and scope of collaboration among researchers is very broad, as only 77 articles are single-authored. There are also only 63 authors of single-author articles. These circumstances indicate that the hotspots and depth of research in statistics with the development of big data are getting deeper and deeper, and most of the topics cannot be completed by a single author at all, and it is necessary to strengthen the cooperation between researchers especially with the integration of cross-disciplinary development of research.

3.2 Distribution of authors' countries

From the statistical data table of the number of articles published by corresponding authors and the statistical chart, it can be seen that the country with the highest number of articles published as corresponding authors is the United States with 524 articles, including 408 articles with all American authors and 116 articles with non-American authors, with an international cooperation rate of 0.221. The country with the second highest number of corresponding author publications is China, with a total of 234 articles. Among them, there are 142 articles with all Chinese authors and 92 articles with non Chinese authors. The international cooperation rate of Chinese authors is 0.357. The third place is Canada with 125 articles, including 84 articles with all Canadian authors and 41 articles with non-Canadian authors, the international cooperation rate of Canadian authors is 0.328. The U.S. is still the center of statistical science research, while China and Canada have gradually become the hot countries in this field in recent years, and are a force to be reckoned with in this field. The top three countries account for nearly one-third of the total number of publications worldwide. The data show that although the total number of publications by Chinese authors is less than half of that of the US, the rate of international collaboration by Chinese authors is the highest. This indicates that Chinese research is increasingly involved in international research and is a force to be reckoned with in international research in related fields. As is shown in Figure 1 and Table 1.

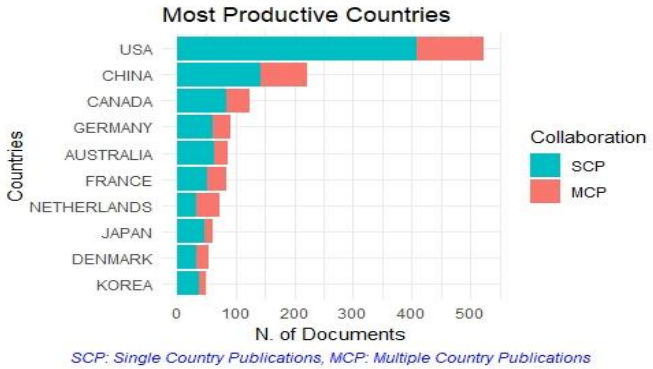

Figure 1: Distribution of Corresponding Authors by Country (Most Productive Countries)

ISSN 2616-5775 Vol. 7, Issue 1: 28-34, DOI: 10.25236/AJCIS.2024.070105

Table 1: Distribution of Corresponding Author's Countries (Corresponding Author's Countries)

No	Country	Articles	SCP	MCP	MCP_Ratio	No	Country	Articles	SCP	MCP	MCP_Ratio
1	USA	524	408	116	0.221	21	NORWAY	17	8	9	0.529
2	CHINA	221	142	79	0.357	22	SINGAPORE	15	8	7	0.467
3	CANADA	125	84	41	0.328	23	SAUDI ARABIA	14	3	11	0.786
4	GERMANY	92	61	31	0.337	24	AUSTRIA	12	7	5	0.417
5	AUSTRALIA	86	64	22	0.256	25	NIGERIA	12	6	6	0.500
6	FRANCE	84	52	32	0.381	26	TURKEY	12	10	2	0.167
7	NETHERLANDS	72	32	40	0.556	27	SOUTH AFRICA	11	4	7	0.636
8	JAPAN	61	48	13	0.213	28	HONGKONG	10	5	5	0.500
9	DENMARK	55	32	23	0.418	29	EGYPT	9	6	3	0.333
10	KOREA	50	38	12	0.240	30	ETHIOPIA	9	8	1	0.111
11	ITALY	48	30	18	0.375	31	FINLAND	8	3	5	0.625
12	UNITEKINGDOM	45	28	17	0.378	32	NEWZEALAND	7	4	3	0.429
13	SPAIN	40	24	16	0.400	33	POLAND	7	4	3	0.429
14	SWEDEN	35	14	21	0.600	34	BANGLADESH	6	5	1	0.167
15	BRAZIL	31	19	12	0.387	35	IRELAND	6	4	2	0.333
16	SWITZERLAND	30	16	14	0.467	36	GREECE	5	5	0	0.000
17	BELGIUM	27	14	13	0.481	37	MALAYSIA	5	3	2	0.400
18	IRAN	25	19	6	0.240	38	MYANMAR	5	0	5	1.000
19	INDIA	21	17	4	0.190	39	PORTUGAL	5	3	2	0.400
20	ISRAEL	17	9	8	0.471	40	MEXICO	4	2	2	0.500

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

3.3 Keyword analysis of published articles

The top 40 keywords in the posting data statistics show that censored data research topics are distributed in various fields of survival analysis, such as: HUMANS, FEMALE, MIDDLE AGED, ADULT, PROPORTIONAL HAZARDS MODELS, SURVIVAL ANALYSIS, RISK FACTORS, etc. According to the statistics of the last five years, anthropology is the most frequently used keyword by authors, with 2038 keywords; feminology is in the second place, with 1156 keywords; masculinity is in the third place, with 1042 keywords; middle-age is in the fourth place, with 768 keywords; and the third place is in the third place, with 768 keywords. In the fourth place is middle-age studies, with 768 keywords; in the fifth place is adult studies, with 677 keywords. Anthropology is the most important research topic in the field of censored data research, with 2,038 keywords in the first place among the total of 2,810 documents. The 10th most frequent keyword was 376 times, which was also the keyword for every 6 documents on average. The statistical results show that the keywords in the top 10 frequencies are the main directions in the field of censored data research in recent years and deserve the attention of researchers. Using this econometric analysis information researchers can more easily find their own research selections, avoid unnecessary literature tracking process, and save valuable research time and money. As is shown in Figure 2 and Table 2.

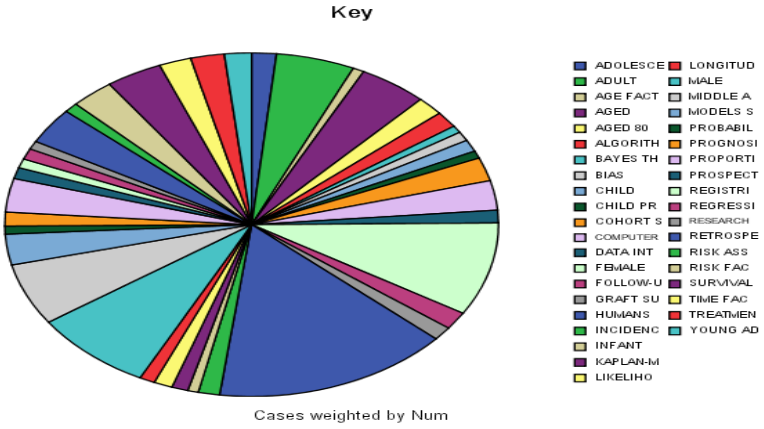

Figure 2: Keyword chart of posted articles

ISSN 2616-5775 Vol. 7, Issue 1: 28-34, DOI: 10.25236/AJCIS.2024.070105

Table 2: Most Relevant Keywords for Published Articles

Sort	Author Keywords(ID)	Articles	Sort	Author Keywords(ID)	Articles
1	HUMANS	2038	21	INCIDENCE	184
2	FEMALE	1156	22	PROGNOS	172
3	MALE	1042	23	GRAFT SURVIVA	169
4	MIDDLEAGED	768	24	LIKELIHOOD FUNCTION	161
5	ADULT	677	25	DATA INTERPRETATION STATISTICAL	157
6	AGED	593	26	CHILD	148
7	RETROSPECTIVE STUDIES	450	27	KAPLAN-MEIER ESTIMATE	141
8	PROPORTIONAL HAZARDSMODELS	415	28	PROSPECTIVE STUDIES	139
9	MODELS STATISTIC	380	29	REGRESSION ANALYSIS	135
10	SURVIVAL ANALYSI	376	30	LONGITUDINAL STUDIES	132
11	RISK FACTORS	374	31	RISK ASSESSMENT	120
12	COMPUTER SIMUL	365	32	BIAS	115
13	COHORT STUDIES	291	33	REGISTRIES	113
14	TREATMENT OUTC	290	34	SURVIVAL RATE	110
15	TIME FACTORS	276	35	RESEARCH DESIG	103
16	YOUNG ADULT	233	36	PROBABILITY	99
17	AGED 80 AND OVER	231	37	AGE FACTORS	97
18	FOLLOW-UPSTUDIES	219	38	CHILD PRESCHOOL	92
19	ADOLESCENT	212	39	INFANT	91
20	ALGORITHMS	201	40	BAYES THEOR	89

3.4 Lokta analysis

Lokta analysis is a statistical analysis method given by the American insurance company statistician Lotka in 1926, which analyzes the productive capacity of scientific and technological workers and their contribution to scientific and technological progress and social development through the statistics of published treatises. The statistical analysis by R software shows that the Lokta parameter corresponding to the subject of this study is $\beta = 2.769, C = 0.685$, the significance p-value of the test is 0.002. In order to facilitate the intuitive cognition of the Lokta analysis, the fit between the theoretical and actual values of the Lokta distribution of this study can be seen from the distribution fit of Figure 3. It shows that the correlation between the transformation of the scientific and technological production capacity of the researchers concerned is relatively obvious. The significance level of the statistical test, P-value 0.002, is much smaller than the significance level, and it is statistically significant that the theoretical and actual values fit well. As is shown in Figure 3.

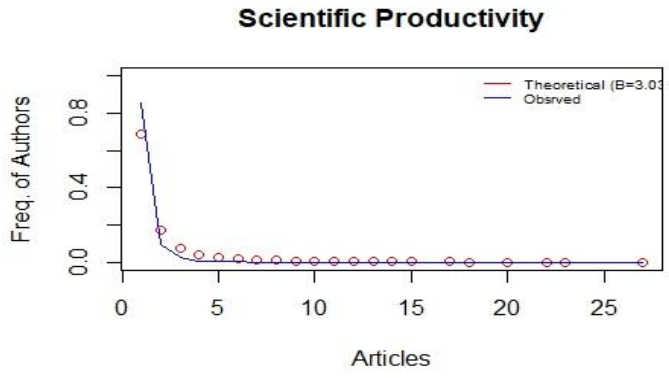

Figure 3: Scientific Productivity Distribution Fitted Plot

3.5 Visual analysis of author collaboration network

Collaborative network analysis is a common research method to study the activity of a specific research field. This method can give the hotspots of research and active research scholars in related fields through graphical representation and author collaboration network analysis. It can point out the more prominent research centers in related fields. The significance of the network is that the larger the nodes of the network, the more collaborators the author has, and the higher the research activity of the

ISSN 2616-5775 Vol. 7, Issue 1: 28-34, DOI: 10.25236/AJCIS.2024.070105

author; the more connected lines indicate the more frequent collaborative research of the authors at both ends of the connection. The network visualization analysis of author collaboration using R software shows that the top 10 authors in terms of number of publications in the last 5 years have more obvious collaborative research experience through the collaborator network visualization in Figure 4. Almost all of the researchers in the top 10 in terms of publication volume have more than 2-3 collaborative researchers. There are 22 collaborative network links in the collaborative network, among which WANG Y., LI Y., and HU Y. are the authors with large collaborative nodes, indicating that these three researchers are more active in collaborative research, and basically form three research collaboration centers, and the research themes of the three research centers are more obvious. The collaboration network diagram of the top 10 authors in terms of publication volume shows that WANG Y. has the largest collaboration node and has collaboration experience with five other authors. The second largest author is ZHANG Y., who has collaborative research experience with four other authors, while SUN J., who has the largest number of publications, has only three co-investigators, a situation that suggests a difference in research style between the researcher and the rest of the researchers. The author with the largest international collaboration is not the author with the highest number of publications. The sum of the research results of the respective research teams of the two authors with the highest collaboration status is even greater than the number of articles published by the collaborative network of the first author. These facts indicate that collaborative research has become an indispensable research tool in the Internet+ era, and it has greatly contributed to the development of global scientific research. As is shown in Figure 4.

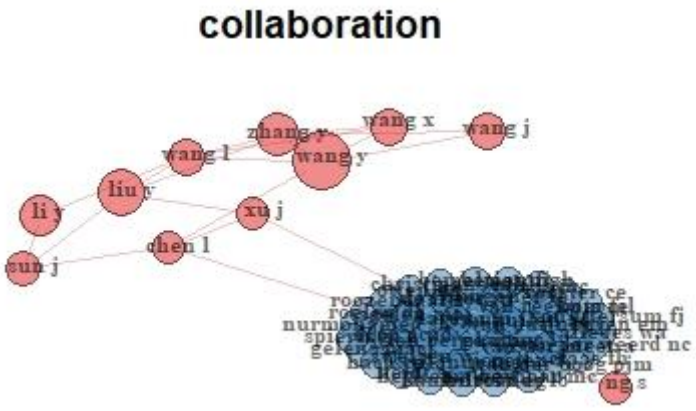

Figure 4: Collaboration Network Visualization

3.6 Semantic map analysis

Strategic coordinate map analysis in semantic map analysis is a two-bit graph built by centripetal degree (X) and density (Y), which is used to measure the development of each topic within a research field and the interaction between topics, where density (Y) is used to evaluate the strength of intra-topic association, indicating the ability of the topic class itself to sustain and develop; centripetal degree (X) is used to evaluate the closeness between one topic class and another topic class, and the larger the centripetal degree indicates the closer the inter-topic connection, the stronger the topic is in a central position in a research field. The greater the centripetal degree, the stronger the connection between themes, and the more central the theme is in a certain research field. From the map of strategic coordinates in Figure 5, it can be seen that in terms of centripetal degree HUMAN and FEMALE keywords each form a research topic class and each becomes the core of that research field. On the other hand, sars-cov, surveys and questionnaires, and progression-free survival have the highest density within the three theme clusters, which indicates that these three theme clusters have strong internal development ability and obvious development trend. This strategy map clearly shows the development trend of thematic clusters as well as internal ones. As is shown in Figure 5.

ISSN 2616-5775 Vol. 7, Issue 1: 28-34, DOI: 10.25236/AJCIS.2024.070105

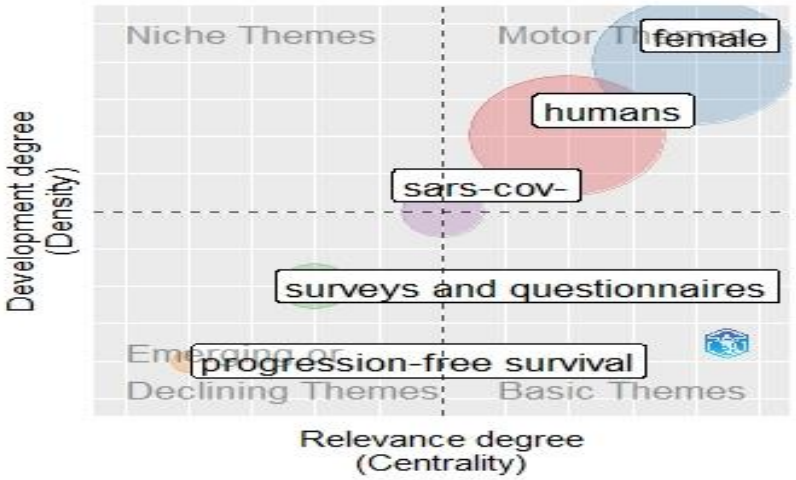

Figure 5: Strategic Coordinate Map

3.7 Keyword cloud analysis

Lexicon analysis is a method to further visualize the presentation of statistical analysis results based on the results of descriptive statistical analysis. Based on the results of the descriptive statistical analysis of the authors' keywords, the word cloud was presented using R software. From Fig. 6, it is clear that the keyword in the first topic position is HUMANS, followed by FEMALE, MIDDLE AGED and other keywords. The word cloud display enables researchers to have a clearer grasp of the research hotspots in the field. These findings are a powerful supplement to the authors' supplementary keyword statistical analysis and provide a valuable reference for relevant researchers to adopt. As is shown in Figure 6.

Figure 6: Keyword word cloud map

4. Conclusion

With the in-depth development of big data, statistical science has also entered the fast track of leapfrog development. But at the same time it also poses more challenging problems for statisticians to solve the urgent difficulties encountered in the development of big data. How to find valuable research information from a large amount of fragmented information in many research fields specifically for each research scholar is a difficulty faced by every researcher, and the scientometric analysis of knowledge given in this paper can provide relevant reference. In this paper, we use the scientometric analysis of knowledge in librarianship to analyze the data of the last five years of censored data in the field of statistics in order to help researchers and related scholars. Through the analysis we find the research hotspots in this field, sort out the development trend in the last 5 years, and discover the international hot research teams as well as the main research scholars. These knowledge mapping

ISSN 2616-5775 Vol. 7, Issue 1: 28-34, DOI: 10.25236/AJCIS.2024.070105

analyses provide a more effective help for researchers to effectively establish their research areas 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), and Zhaoqing College Quality Project and Teaching Reform Project (zlgc202112).

References

[1] Chu Jiewang, Sun Xiaoning. A literature dosage analysis of the current status of knowledge management research in domestic libraries [J]. Library Theory and Practice, 2012(9): 21-26.
[2] Hou Haiyan, Liu Zeyuan, Luan Chunjuan. A knowledge graph-based econometric analysis of international research frontiers in scientometrics [J]. Research Management, 2009, 30 (1):164-170.
[3] Jin Bihui, Zhang Jianggong, Chen Dingquan, et al. Development of the Chinese scientometric indicators (CSI) [J]. Scientometrics, 2002, 54 (1):145-154.
[4] Liang Limin. Scientometrics and informetrics: A world perspective on China [J]. Scientific Research Management, 2000, 21(3):95-101.
[5] Chen Dinquan, Li Shuai. Key Issues for Investment in Library Technology Infrastructure [J]. Library Research, 2023, 12 (6):123-125.
[6] Liang, G. A review of domestic bibliometrics[J]. Science and Technology Literature Information Management, 2013 (4):58-59, 62.
[7] Li Fengzhi. Definition of citation analysis method and its role [J]. Science and Technology Information, 2015(10): 25-253.
[8] Li Weichao, Guo Jun, Li Jingyu. Comparison of domestic and international research hotspots in library intelligence [J]. New Century Library, 2021(3): 12-17.
[9] Lin Lili, Ma Xiufeng. Analysis of domestic library intelligence research theme discovery and evolution based on LDA model [J]. Intelligence Science, 2019(12): 87-92.
[10] Hou Jianhua, Cai Yangxiu, Zhou Lijuan. Analysis of frontier themes of international research in the field of library intelligence and their evolutionary trends[J]. Library Intelligence Work, 2016(13): 82-90.
[11] Ma F, Yang S-L. Comparative analysis of the thematic content of domestic and foreign library intelligence research [J]. Intelligence Science, 2015(9): 140-145.
[12] Teng Guangqing, Mou Dongmei, Ren Jing. Research on the application of foreign social network analysis in the field of bibliometrics[J]. Intelligence Data Work, 2014(1): 47-51.
[13] Si Li, et al. Comparative analysis of research hotspots in domestic and foreign library intelligence and archives management disciplines from 2014-2018 [J]. Books and Intelligence, 2020(1): 75-82.
[14] Su XN, Xia LX. Analysis of digital library research topic areas in China from 2000 to 2009 - based on CSSCI keyword statistics [J]. Chinese Journal of Library Science, 2011, 37(7):60-69.
[15] Shi Yanqing, Sun Jianjun. Analysis of research theme characteristics of China's library intelligence disciplines in the international arena [J]. Library and intelligence work, 2018(7): 66-76.
[16] Wang W, Wang LW, Zhu H. Analysis of international research frontiers and hotspots in informetrics [J]. Journal of Medical Informatics, 2010, 31 (2): 1-4.