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 adopts bibliometric theory and uses the R software for data analysis to investigate research literature on missing data retrieved from the PubMed database between 2017 and 2022. We first extract basic bibliographic information such as title, abstract, authors, affiliations, journals and citation records from PubMed, and perform data cleaning, deduplication and standardization before subsequent computation. Co-author analysis and institutional co-occurrence analysis are applied to reveal the main collaborative groups and geographical distribution of active research teams. Meanwhile, keyword co-occurrence analysis and burst detection are implemented to identify frontier topics and track the shift of research focus during this period. This study aims to sort out the evolution of research themes in this field over the past five years and provide references for researchers in relevant disciplines. The findings can help new entrants quickly grasp the overall framework of missing data research in biomedicine, identify influential scholars and high-impact journals, and capture emerging directions such as multiple imputation, sensitivity analysis and missing not at random mechanisms. Furthermore, it offers clues for scholars to seek potential partners and design follow-up research plans.*

Keywords: *library intelligence, bibliometric 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.

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

2. Preparation of data set

Missing data is a pervasive challenge in clinical trials, epidemiological surveys and biomedical observational studies. Improper handling of missing values may lead to biased effect estimation, inflated type I error rates and unreliable research conclusions, which has gradually attracted widespread attention from methodologists and clinical researchers. Although multiple review articles have summarized statistical methods for missing data, few studies systematically map the intellectual landscape of this research domain from a macroscopic perspective. Bibliometrics serves as a powerful quantitative tool to objectively describe the distribution characteristics, collaborative networks and dynamic trends of academic publications, which can compensate for the subjectivity of traditional narrative reviews. Knowledge measurement and visualization analyses are carried out from multiple perspectives, including publication volume, citations, collaboration patterns, core authors and research hotspots. PubMed, one of the most authoritative and mainstream international databases for biomedical research, was chosen as the data source in this study. We adopted the built-in professional search tool of PubMed and set "censored data" as the search keyword, with the time window ranging from January 1, 2017 to May 1, 2022. A total of 3733 bibliographic records were initially retrieved. After excluding books and other document types that failed to meet the inclusion criteria for subsequent bibliometric analysis, 2810 valid records were retained for further processing. The open-source statistical software R was used to conduct statistical analysis, visualization and bibliometric analysis on the final dataset of 2810 publications. Multiple fields of bibliographic information were extracted and analysed, including supplementary keywords (ID), author keywords (DE), abstracts (AB), authors' publication productivity, as well as the country and institutional affiliations of contributors. This study intends to clarify the status quo, research hotspots and evolving trends of censored data research within survival analysis. It is expected to offer guidance for relevant researchers in this domain and help them design research topics with better scientific rigor and forward-looking perspectives.

3. Bibliometric analysis

3.1 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.

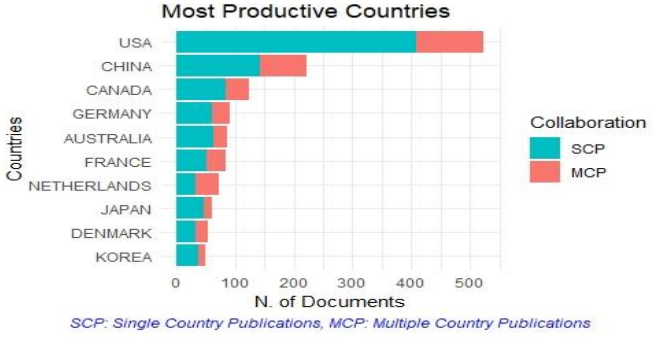


Figure 1: Distribution of Corresponding Authors by Country (Most Productive Countries)

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

3.2 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.

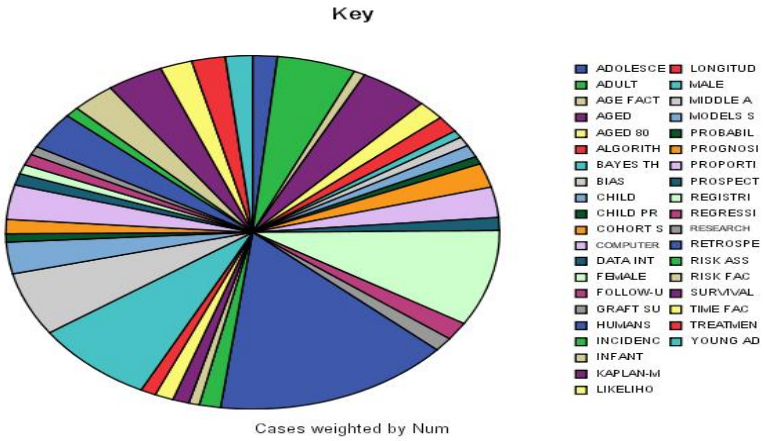


Figure 2: Keyword chart of posted articles

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

3.3 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.

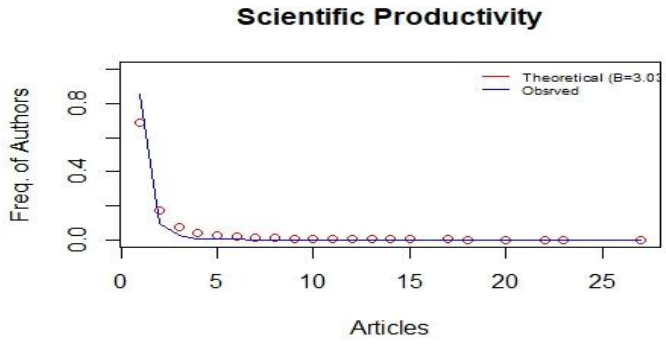


Figure 3: Scientific Productivity Distribution Fitted Plot

3.4 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 author; the more connected lines indicate the more frequent collaborative research of the authors at

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

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.

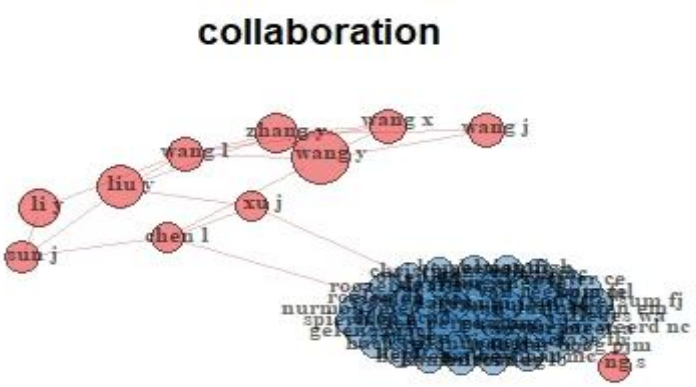


Figure 4: Collaboration Network Visualization

3.5 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 5: Keyword word cloud map

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

4. Conclusion

This study took the PubMed database as the data source and demonstrated the cross-application of information measurement methods in the field of Biostatistics literature. Through statistical analyses of the nationality distribution of corresponding authors, the productivity of countries and institutions, the distribution of core journals, and the keyword co-occurrence and word cloud maps, the study reveals the overall landscape, research hotspots, and cooperative networks of publications in this field. The results show that information measurement methods such as bibliometric analysis, citation analysis, and knowledge mapping, which originated in library and information science, can be effectively transplanted to the biomedical domain, providing quantitative evidence for identifying research frontiers, evaluating disciplinary influence, and optimizing scientific resource allocation. These analytical tools can complement traditional narrative reviews and help researchers objectively capture implicit academic links that are difficult to identify through manual reading alone. In future research, the following directions deserve further exploration: first, constructing multi-database fusion measurement models that integrate PubMed with Web of Science, Scopus, and other sources to reduce the bias of a single data source; and second, developing dynamic, real-time knowledge measurement and early-warning systems so that information measurement methods can move from static retrospective analysis toward predictive decision support. Moreover, combined with text mining and Bibliometric analysis, fine-grained topic evolution tracking can be realized to dig deeper into the internal logical connections among different research branches.

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] Dai Guoqiang, Zhu Ning, Liu Gang. Analysis of the Reasons for the Decline in the Borrowing Volume of Paper Books in University Libraries and Countermeasures: A Case Study of the Library of Yibin University [J]. Office Automation, 2019, 24(2): 56-57, 21.

[2] Hou Jianhua, Yang Xiucai, 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.

[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, L. M. Scientometrics and Informetrics: A World Perspective on China [J]. Scientific Research Management, 2000, 21(3): 95-101.

[5] Liang, L. M. The Development of Scientometrics and Informetrics: A Comparative Study Between Mainland China and Taiwan [J]. Library and Information Service, 2012, 56(18): 5-12

[6] Liu Jingkun. Research on College Students' Reading Preferences Based on Learning Analytics: A Case Study of School J, University X [D]. Wuhan: Huazhong University of Science and Technology, 2018.

[7] Ma Feng, Yang Siluo. Comparative Analysis of the Thematic Content of Domestic and Foreign Library and Information Science Research [J]. Information Science, 2015, 33(9): 140-145.

[8] Pang Bo. Research on Reading Promotion Services in University Libraries [D]. Harbin: Heilongjiang University, 2017: 10.

[9] Qi Wenjuan, Sun Jing, Hu Jingjun. Research on the Path of Cultural Education in Universities in the New Era: Based on the Analysis of College Students' Book Borrowing Behavior [J]. Times Finance, 2019(5): 201, 227.

[10] Shi Junjie. Statistical Analysis of the Collection and Borrowing Status of Professional Libraries: A Case Study of the Dunhuang Academy [J]. Gansu Science and Technology, 2018, 34(1): 57-60.

[11] Shi Yanqing, Sun Jianjun. Analysis on the Theme Structure and Features of Chinese Library and Information Science in International Studies [J]. Library and Information Service, 2018, 62(7): 66-76.

[12] Tang Risheng. Reasons for the Low Borrowing Rate of Paper Books in University Libraries and Improvement Measures [J]. Journal of Liaoning Agricultural Vocational and Technical College, 2018(6): 54-55.

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

[13] Teng Guangqing, Mou Dongmei, Ren Jing. Research on the Application of Foreign Social Network Analysis in the Field of Bibliometrics [J]. Information and Documentation Services, 2014, 35(1): 47-51.
[14] Wang Jing. Information Metrological Analysis Based on the Data of PubMed Database from 2017 to 2022 [J]. Inner Mongolia Science Technology & Economy, 2024(16): 138-142.
[15] Wang Wei, Wang Liwei, Zhu Hong. Analysis of International Research Frontiers and Hotspots in Informetrics [J]. Journal of Medical Informatics, 2010, 31(2): 1-4.
[16] Wang Ying. Statistics and Analysis Based on the Circulation Data of University Libraries [J]. Journal of Changchun Normal University, 2017, 36(10): 171-173, 114