The rise of large language models in the medical field: A bibliometric analysis

Wenhao Qi
School of Nursing, Hangzhou Normal University, Hangzhou,Zhejiang,China,1535569 3529@163.com

Shihua Cao†
School of Nursing, Hangzhou Normal University, Hangzhou,Zhejiang,China,csh@hznu.edu.cn

Bin Wang
School of Nursing, Hangzhou Normal University, Hangzhou,Zhejiang,China,2023112 026046@stu.hznu.edu.cn

Xiaohong Zhu
School of Nursing, Hangzhou Normal University, Hangzhou,Zhejiang,China, 2023112026025@stu.hznu.edu.cn

Chaoqun Dong
School of Nursing, Hangzhou Normal University, Hangzhou,Zhejiang,China, 1915284824@qq.com

Danni He
School of Nursing, Hangzhou Normal University, Hangzhou,Zhejiang,China, 2021112012182@stu.hznu.edu.cn

Yanfei Chen
School of Nursing, Hangzhou Normal University, Hangzhou,Zhejiang,China, 2021112012203@stu.hznu.edu.cn

Yankai Shi
School of Nursing, Hangzhou Normal University, Hangzhou,Zhejiang,China, shiyankai@stu.hznu.edu.cn

BingSheng Wang
School of Nursing, Hangzhou Normal University, Hangzhou,Zhejiang,China, 2022112024058@stu.hznu.edu.cn

***Abstract*—Large language models, through pre-training on massive textual data, have become powerful tools for processing medical texts, extracting key information, and insights. **Objective:** This study, through literature data mining, analyzes and evaluates the key contributions and development scale of large language models in the medical field. **Methods:** Literature search was conducted using the Web of Science core database, covering publications up to December 2023. Bibliometrics and content analysis were utilized to explore the current status, hotspots, and trends of research. **Results:** A total of 583 papers were included in the analysis, published across 328 journals, involving 1275 institutions from 70 countries, and 3369 authors. The field saw rapid development within 2023, with a significant increase in publications, accounting for 71.7% (418/583) of all papers. BERT and ChatGPT are the most used large language models in the medical field. Their primary applications are in medical education, medical literature processing, and electronic health records. **Conclusion:** Large language models are rapidly evolving in the medical field, especially in the United States, China, and the United Kingdom. Collaboration among different author groups should continue to be strengthened. Their application in disease will be a major research hotspot and trend in the future, but challenges such as privacy and ethical issues that accompany their development must be cautiously addressed.**

Keywords—Large Language Models, LLMS, Medical, BERT, ChatGPT

I. INTRODUCTION

With the rapid advancement of artificial intelligence (AI) technology, Large Language Models (LLMs) and their implementation in various applications, especially chatbots like ChatGPT, have attracted widespread attention [1]. Since its release in November 2022, ChatGPT has garnered over a hundred million users in just two months, significantly marking the popularity and potential applications of LLM technology [2]. LLMs, built on the deep learning Transformer architecture, are capable of processing, generating, and understanding human-like natural language texts [3]. These models are trained by analyzing and learning from billions of words available on the internet (including articles, books, etc.), thereby demonstrating unprecedented complexity and efficiency in natural language processing (NLP) tasks [4].

979-8-3503-6345-6/24/$31.00 ©2024 IEEE
DOI 10.1109/FMSys62467.2024.00014

Particularly in the medical field, the potential applications of LLMs have sparked tremendous interest. This interest stems from LLMs' enormous potential in handling and understanding biomedical NLP tasks, with the expectation that they will play a key role in reshaping information processing and knowledge management within the healthcare sector [5].

Scientometrics is extremely important in assessing the scientific research of a specific discipline or area [6]. Currently, while many studies explore the specific use of large language models in the medical field, there is a lack of bibliometric analysis to systematically summarize and evaluate the scale and quality of existing research. Therefore, through scientometric analysis, we have revealed the current state, hotspots, and trends in this field, providing recommendations for in-depth research in this area.

II. Method

A. Data and search strategy

The Web of Science Core Collection is considered by many researchers to be the central resource for interdisciplinary academic research [7]. Therefore, we chose it as our search database. After conducting an online search in the Web of Science Core Collection, removing duplicates, excluding papers unrelated to the topic, and papers of types such as information, conferences, and news, a total of 583 papers were included. The search formula and inclusion process are shown in Figure 1.

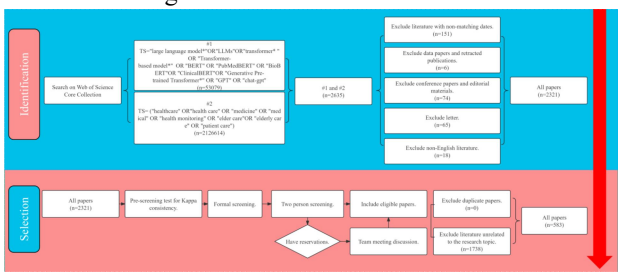

Fig. 1. Literature search and inclusion process flowchart.

B. Data analysis

- Analysis tool

VOSviewer (1.6.19): A free Java document mapping software capable of constructing various visual networks [8], exploring network structures.

CiteSpace (5.7.R5): A Java-based scientometric tool used for analyzing aspects such as funding sponsorships in research studies [9].

Bibliometrix: An R-language-based tool equipped with numerous functions for processing and analyzing literature data [10].

Gephi (0.10.1): A highly autonomous software for visualizing social networks.

To display regional visualizations and enhance the readability of the map, we introduced ScimagoGraphica and Pajek64 Portable (5.18) for layout purposes.

- Statistical Analysis

When mapping keywords, organizations, and author nodes using VOSviewer (1.6.19) and Bibliometrix, a certain threshold was set to enhance the clarity of nodes and connectors, thereby improving the readability of the graphics. The Origin2021 software, in conjunction with Excel, was used for the analysis of annual publications and journal publications, employing the least squares method for graph plotting as well as polynomial fitting and linear fitting of publications, with the best fit trend line model selected based on the R^2 value. Core author collaboration mapping utilized Price's law calculation [11] to determine the minimum number of articles. Bradford's Law [12] was followed to ascertain the number of core journals. For author keywords, the bibliometrix package in R was used to integrate synonyms before conducting co-occurrence analysis, thus ensuring the uniformity and accuracy of keywords. For instance, "NLP" was merged into Natural Language Processing.

III. Result

A. The Annual trends of publications

The changes in the number of publications reflect the development dynamics of the field. Research related to this field began as early as 2019, with both the number of publications and citations peaking in 2023, totaling 418 papers and 1969 citations respectively. According to the trend line description, the publication activity in this field is on an upward trajectory, as shown in Figure 2.

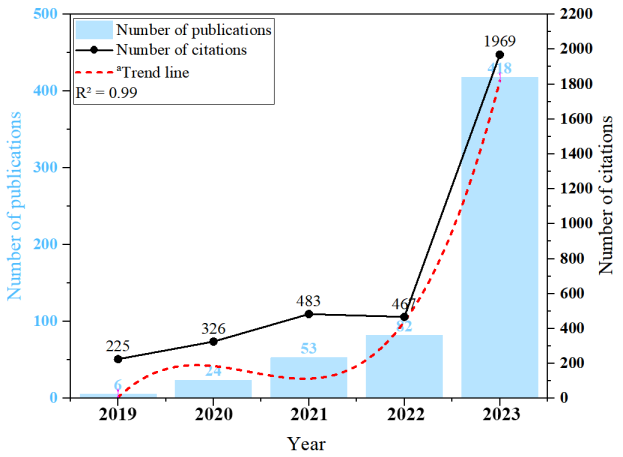

Fig. 2. Publication trends for large language models in the medical field from 2019 to 2023.

*Trend line:y = 24.667x3 - 176.57x2 + 397.76x - 244.4

B. Author analysis

The number of participants from researchers reflects the level of interest in the application of large language models in the medical field. In this domain, a total of 3,369 authors contributed to the publication of 3,770 studies. Among them, the authors with the most publications were Mottrie, Aramaki, Eiji, and Seth, Ishith (6 papers each). According to Price's Law, the core author threshold is approximately 2 papers. 222 individuals (6.4%) met this criterion, contributing to 505 papers (13.4%), which did not meet Price's Law (>50%) standard. In the collaboration network graph, overall, the co-occurrence network among core authors is relatively independent, with fewer connections, exhibiting a phenomenon of high cohesion and low coupling, as depicted in Figure 3.

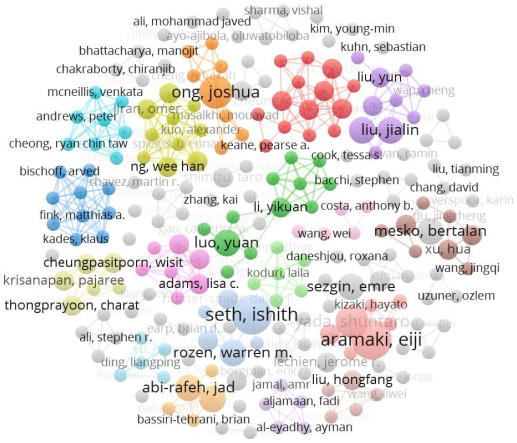

Fig. 3. The collaboration network graph of core authors.

C. Institutions analysis

Institutional analysis has unveiled the structural characteristics of organizations engaged in research within the dementia biomarkers domain. A total of 1,275 institutions have published related studies. Among these, Stanford University leads with 22 publications, followed by Mayo Clinic with 14 publications, and National University of Singapore with 13 publications, ranking third. It's noteworthy that among the top ten producing institutions (with the minimum publication count being eight, allowing ties, thus including 11 institutions in total), all except Mayo Clinic are university institutions, and four are from the United States, as shown in Table 1. By selecting the top 100 institutions by publication volume (with the minimum publication count being three), a collaboration network was established, displaying 382 collaborative connections, as seen in Figure 4.

TABLE I. TOP 10 Institutions in medical field output by large language models.

Rank	Organization	Output （N=1907) ,n%	Country
1	Stanford University	22(1.2)	United States
2	Mayo Clinic	14(0.7)	United States
3	National University of Singapore	13(0.7)	Singapore
4	University of Michigan	11(0.6)	United States
5	Monash University	10(0.5)	Australia
6	Harvard Medical School	9(0.5)	United States
7	Peking University	8(0.4)	China
7	Sichuan University	8(0.4)	China
7	University of Oxford	8(0.4)	United Kingdom
7	University of Toronto	8(0.4)	Canada
7	Yale University	8(0.4)	United States

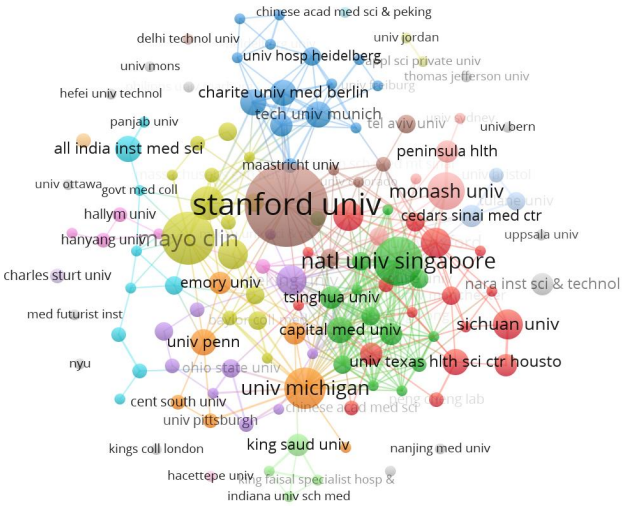

Fig. 4. Collaborative network of research institutions in the medical field on large language models.

D. Journal analysis

Journal analysis reveals the structure of disciplines and the characteristics of journals. A total of 328 journals have published relevant articles. Following Bradford's Law, we identified 17 core journals in this field, contributing a total of 221 research articles (38.0%), as shown in Figure 5.

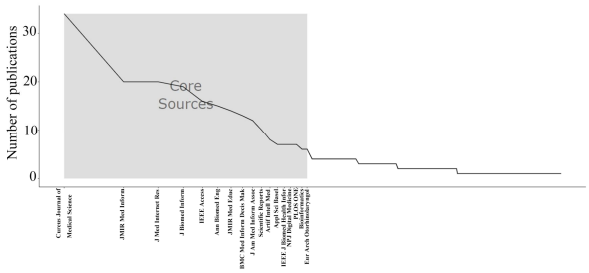

Fig. 5. Core journal output distribution graph.

E. National/Regional analysis

The number of countries/regions involved in related research reflects the global interest in the topic. A total of 70 countries or regions have published research, with the United States, China, and the United Kingdom leading globally with 229, 123, and 61 publications, respectively. The top ten countries combined contribute 646 research papers, accounting for 70.0% of the total, as shown in Table 2. An international collaboration map was drawn based on the cooperation relationships between countries, with thicker lines indicating more collaboration. Many countries/regions have engaged in relevant cooperation, among which the United States, China, and the United Kingdom have the most frequent collaborations, as seen in Figure 6.

TABLE II. Top 10 Countries or regions in medical field output by large language models.

Rank	Country	Output(N=923),n%
1	United States	229(25.0)
2	China	123(13.3)
3	United kingdom	61(6.6)
4	Germany	47(5.1)
5	India	42(4.6)
6	Canada	34(3.7)
7	Australia	33(3.6)
8	South korea	29(3.1)
9	Italy	24(2.6)
10	Japan	24(2.6)

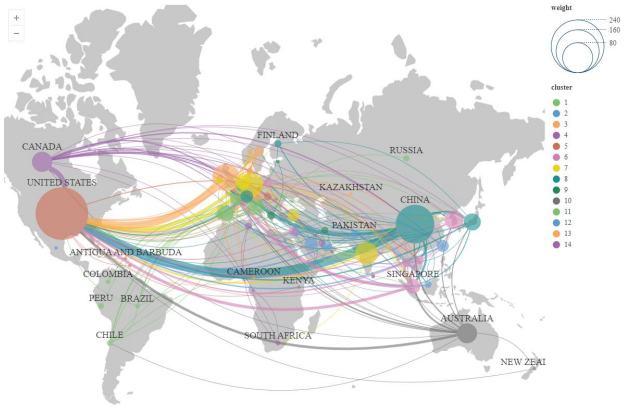

Fig. 6. Global application of large language models in the medical field and international cooperation map by country.

F. Grant analysis

The funding analysis reveals the support provided by various countries for research. In total, 339 distinct funds sponsored research in this area 425 times. Among these, the National Natural Science Foundation of China, a Chinese fund, supported 28 related research projects, ranking first. It is noteworthy that five of the top ten funding entities by number of sponsorships are from the United States. This information reflects, to a certain extent, the financial commitment and emphasis placed on this field by both China and the United States, as shown in Table 3.

TABLE III. Top 10 Funding organizations by number of grants in the application of large language models in medicine.

Rank	Grant	Frequency(N=425),n%	Country
1	NSFC	28(6.6)	China
2	NIH	11(2.6)	United States
3	NKRDPC	11(2.6)	China
4	European Union	5(1.2)	International Organization
5	NSF	4(0.9)	United States
6	AHA	3(0.7)	United States
7	NCI	3(0.7)	United States
8	NIA	3(0.7)	United States
9	NMRC	3(0.7)	Singapore
10	NRF	3(0.7)	South korea

G. Keyword analysis

After merging synonymous keywords, a total of 1,674 unique keywords were obtained, with a combined frequency of 3,301 occurrences. Among them, "Artificial intelligence" appeared 224 times (6.8%), "ChatGPT"

appeared 207 times (6.3%), and "Large language model" appeared 136 times (4.1%). A time-heat map was generated for the top 20 highest-frequency keywords, showing that the frequency of occurrences of the BERT model has been consistently high over the past five years. On the other hand, occurrences of ChatGPT began to gradually increase in 2021. Ultimately, these high-frequency keywords all exhibited extremely high frequencies in 2023, as shown in Figure 7A. Figure 7B presents a co-occurrence network of keywords, encompassing specific diseases such as COVID-19, specific large language models like BERT, and specific tasks such as diagnosis.

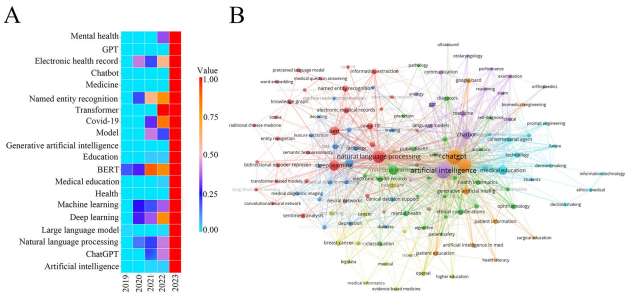

Fig. 7. (A) Heatmap of high-frequency keywords. (B) Co-occurrence network of keywords related to large language models in the medical field.

H. Specific applications of large language models in the medical field

We constructed a co-occurrence network between the specific models used and the specific medical disciplines or diseases in 583 studies. Currently, 9 major large language models are utilized in the medical field, applied to 69 different tasks, specific diseases, clinical departments, or sub-disciplines. Among these, ChatGPT and BERT models are far ahead in terms of usage, with 203 and 135 instances respectively.

In terms of specific diseases, large language models have been applied in some diseases, especially in COVID-19 and oncological diseases (such as lung cancer, breast cancer). In specific medical sub-disciplines, ophthalmology as well as radiology and imaging have seen more frequent use. Within these specific sub-disciplines, large language models are more often used in medical education. In specific medical text tasks, large language models are mainly used for processing electronic health records and medical literature. See Figure 8.

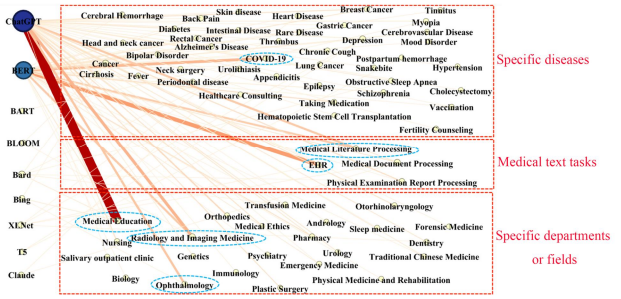

Fig. 8. Connection between large language models and specific medical tasks.

IV. Discussion

A. Current status of research

Through this study, we have identified some specific situations in the field. In terms of publication quantity, the number of publications between 2019 and 2023 increased by 69.7 times. The involvement of 3,369 researchers in publishing research underscores the significant attention paid by numerous researchers to the application of large language models in the medical field in recent years. However, there is limited connection among various author groups, indicating the need for enhanced collaboration. In terms of geographical distribution, 70 countries or regions globally have reported relevant studies, forming a closely-knit cooperation network among these nations or regions. In addition to developed Western countries like the United States and the United Kingdom, some developing countries such as China have also made significant contributions to research in this field. In institutional analysis, we found that Stanford University and the Mayo Clinic have relatively high output in this field, reflecting the attention paid to this topic by strong research universities and medical institutions. Cooperation among institutions also shows a widespread trend. In journal analysis, 328 journals have published related research, with 15 out of 17 core journals being SCI-indexed, indicating the attention of numerous high-quality journals to this topic. Overall, the application of large language models in medicine is currently undergoing rapid development and demonstrates significant growth potential.However, during our analysis of funding, we found that not all studies received financial support. The funds with high instances of support mainly come from the USA and China, particularly the National Natural Science Foundation of China (NSFC). As more research in this field is expected in the future, scientific research management institutions in various countries should pay attention to this issue, providing the necessary financial support to a wide range of researchers to facilitate the development of related studies.

B. Research hotspots

Through co-occurrence analysis and content mining, we have discovered that the primary large language models currently being used in the medical field are ChatGPT and BERT models. Although ChatGPT gained popularity later than BERT, its usage has surpassed BERT, making it the most widely used large language model in the medical field today. Among these models, the GPT-3.5 model of ChatGPT is the most frequently utilized version, which may be attributed to its higher performance and more accurate text generation capabilities compared to its predecessors. This is particularly important in handling complex medical literature, case reports, and medical consultations. Additionally, compared to GPT-4.0, its free nature makes it more accessible to researchers. On the other hand, BERT has more variants and improved versions, such as Sentence-BERT, PubMedBERT, and RoBERTa, all of which are based on BERT and have considerable usage. These large language models have broad applicability but are primarily used in specific domains such as medical education, radiology and imaging, electronic health records, medical literature processing, COVID-19, and ophthalmology. Particularly, ChatGPT is predominantly utilized in medical education, far exceeding its usage in other tasks or diseases, while BERT seems to be more inclined towards handling electronic medical record content.

C. Research trends

Currently, large language models have been researched in certain specific diseases or specific medical sub-disciplines, but there are still many gaps. As research deepens, large language models will be used in the study of more different diseases. Moreover, with the emergence of more large language models, comparative studies between various models may gradually increase in the future.

While the application of large language models (LLMs) in the medical field brings immense potential, it also raises a series of ethical issues. These issues mainly involve data privacy, information accuracy, accountability, bias, and inequality [13]. However, compared to the deployment of LLMs in specific diseases such as breast cancer, heart disease, and skin diseases, there is relatively less research conducted on medical ethics aspects. Whether it's diagnosing or assisting in the treatment of specific diseases through electronic health records or real patient dialogues, LLM training requires a large amount of data, much of which involves sensitive personal health information. Ensuring that the collection, processing, and storage of this data comply with privacy protection laws and ethical standards is an important issue. Additionally, it is necessary to ensure that the outputs of the models do not disclose personal information.

V. SUMMARY

We analyzed the relevant literature on the use of large language models in the medical field in the Web of Science core database and visualized the findings. We found that this field is in a rapid development stage, with an increasing trend in the number of published papers, indicating that research topics will continue to remain hot. Many research topics in this field are extensive and diverse, with medical education and medical literature processing being popular areas of study. Ethical and privacy issues are expected to be major research topics in the future. Additionally, this study only included relevant literature from the Web of Science core dataset, which may result in the omission of some potential articles. There are some limitations, and the next step should involve expanding the search scope, refining research topics, and conducting in-depth studies.

ACKNOWLEDGMENTS

The Grant sponsors of this study are the key Research Project of Laboratory work in Colleges and Universities of Zhejiang Province (ZD202202), the Zhejiang Traditional Chinese Medicine Inheritance and Innovation Project(2023ZX0950), the Medical and Health Technology Plan of Zhejiang Province (2022507651).

REFERENCES

[1] Shah, N. H., Entwistle, D. and Pfeffer, M. A, 2023. Creation and Adoption of Large Language Models in Medicine. Jama, 330, 9 (Sep 5 2023), 866-869.DOI: 10.1001/jama.2023.14217.

[2] Li, R., Kumar, A. and Chen, J. H, 2023. How Chatbots and Large Language Model Artificial Intelligence Systems Will Reshape Modern Medicine: Fountain of Creativity or Pandora's Box? JAMA internal medicine, 183, 6 (Jun 1 2023), 596-597.DOI: 10.1001/jamainternmed.2023.1835.

[3] Cascella, M., Semeraro, F., Montomoli, J., Bellini, V., Piazza, O. and Bignami, E, 2024. The Breakthrough of Large Language Models Release for Medical Applications: 1-Year Timeline and Perspectives. J Med Syst, 48, 1 (Feb 17 2024), 22.DOI: 10.1007/s10916-024-02045-3.

[4] Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F. and Ting, D. S. W, 2023. Large language models in medicine. Nat Med, 29, 8 (Aug 2023), 1930-1940.DOI: 10.1038/s41591-023-02448-8.

[5] Li, H., Moon, J. T., Purkayastha, S., Celi, L. A., Trivedi, H. and Gichoya, J. W, 2023. Ethics of large language models in medicine and medical research. The Lancet. Digital health, 5, 6 (Jun 2023), e333-e335.DOI: 10.1016/S2589-7500(23)00083-3.

[6] Bordons, M. and Zulueta, M. A, 1999. [Evaluation of the scientific activity through bibliometric indices]. Revista espanola de cardiologia, 52, 10 (Oct 1999), 790-800.DOI: 10.1016/s0300-8932(99)75008-6.

[7] Wu, C. C., Huang, C. W., Wang, Y. C., Islam, M. M., Kung, W. M., Weng, Y. C. and Su, C. H, 2022. mHealth Research for Weight Loss, Physical Activity, and Sedentary Behavior: Bibliometric Analysis. J Med Internet Res, 24, 6 (Jun 8 2022), e35747.DOI: 10.2196/35747.

[8] van Eck, N. J. and Waltman, L, 2010. Software survey: VOSviewer, a computer program for bibliometric mapping. Scientometrics, 84, 2 (Aug 2010), 523-538.DOI: 10.1007/s11192-009-0146-3.

[9] Osinska, V. and Klimas, R, 2021. Mapping science: tools for bibliometric and altmetric studies. Inf. Res., 26, 4 (Dec 2021), 20.DOI: 10.1002/asi.20317.

[10] Aria, M. and Cuccurullo, C, 2017. bibliometrix: An R-tool for comprehensive science mapping analysis. J. Informetr., 11, 4 (Nov 2017), 959-975.DOI: 10.1016/j.joi.2017.08.007.

[11] Price, D. J, 1965. NETWORKS OF SCIENTIFIC PAPERS. Science (New York, N.Y.), 149, 3683 (1965-Jul-30 1965), 510-515.DOI: 10.1126/science.149.3683.510.

[12] Weinstock, M. Bradford's Law, 1971. Nature, 233, 5319 (1971-Oct-08 1971), 434-434.DOI: 10.1038/233434a0.

[13] Akinci D'Antonoli, T., Stanzione, A., Bluethgen, C., Vernuccio, F., Ugga, L., Klontzas, M. E., Cuocolo, R., Cannella, R. and Koçak, B, 2023. Large language models in radiology: fundamentals, applications, ethical considerations, risks, and future directions. *Diagnostic and interventional radiology (Ankara, Turkey)* (Oct 3 2023).DOI: 10.4274/dir.2023.232417.