ORIGINAL ARTICLE

Medical Journal of Western Black Sea
Batı Karadeniz Tıp Dergisi

Med J West Black Sea 2026; Early View
DOI: 10.29058/mjwbs.1733561

Applications and research trends of large language models in healthcare: A bibliometric analysis

Sağlık hizmetlerinde büyük dil modellerinin uygulamaları ve araştırma eğilimleri: Bibliyometrik bir analiz

Hakan Öztürk[1], **Elvan Hayat**[2]

[1]Department of Biostatistics, Aydın Adnan Menderes University, Faculty of Medicine, Aydın, Türkiye
[2]Department of Econometrics, Aydın Adnan Menderes University, Faculty of Political Sciences, Aydın, Türkiye

Cite this article as: *Öztürk H and Hayat E. Applications and research trends of large language models in healthcare: a bibliometric analysis. Med J West Black Sea. 2026; Early View.*

ABSTRACT

Aim: This study aims to reveal the current scientific studies and research trends on the use of Large Language Models (LLM) in the field of health by bibliometric analysis.

Material and Methods: Data were sourced from open-access clinical research articles published between 2023 and 2025, extracted from the Web of Science database. A total of 173 articles meeting the inclusion criteria were analyzed using the "Bibliometrix" R package and the "Biblioshiny" interface, focusing on publication trends, co-authorship networks, keyword co-occurrences, and thematic mapping.

Results: Results revealed a substantial rise in academic interest in LLMs within healthcare, with significant contributions from researchers in the United States, China, and various European countries. Prominent keywords such as "Artificial Intelligence," "ChatGPT," and "Medical Education" indicate that LLMs are increasingly explored for educational and clinical support applications. Thematic analysis identified emerging research areas including "performance," "health disparities," and "ethical challenges." While LLMs offer considerable opportunities in healthcare, they also present notable risks like data privacy issues, misinformation, and ethical oversight gaps.

Conclusion: In conclusion, this study provides a comprehensive overview of LLM applications in healthcare, highlighting key research themes and identifying gaps within the current literature.

Keywords: Bibliometric analysis, ChatGPT, generative artificial intelligence, large language models, medicine

ÖZ

Amaç: Bu çalışma, sağlık alanında Büyük Dil Modelleri'nin (LLM) kullanımına ilişkin mevcut durumu ve araştırma eğilimlerini bibliyometrik analiz yöntemiyle ortaya koymayı amaçlamaktadır.

Gereç ve Yöntemler: Veriler, 2023-2025 yılları arasında yayımlanmış açık erişimli klinik araştırma makaleleri arasından seçilerek Web of Science veri tabanından elde edilmiştir. Dahil etme kriterlerini karşılayan 173 makale, R programında "Bibliometrix" paketi ve "Biblioshiny" arayüzü kullanılarak analiz edilmiştir. Yayın eğilimleri, ortak yazarlık ağları, anahtar kelime eş-oluşumları ve tematik haritalama gibi temel bibliyometrik göstergeler incelenmiştir.

Bulgular: Bulgular, sağlık alanında LLM'lere yönelik bilimsel ilginin belirgin şekilde arttığını ve ABD, Çin ve bazı Avrupa ülkelerinin bu alandaki araştırmalara öncülük ettiğini göstermektedir. "Yapay Zekâ", "ChatGPT" ve "Tıp Eğitimi" gibi anahtar kelimelerin öne çıkması, LLM'lerin eğitim ve klinik destek sistemlerinde giderek daha fazla araştırıldığını ortaya koymaktadır. Tematik analizde, "performans", "sağlık eşitsizlikleri" ve "etik zorluklar" gibi konuların yükselen araştırma alanları olduğu görülmüştür. LLM'ler sağlık hizmetlerinde önemli fırsatlar sunarken, veri gizliliği, yanlış bilgi üretimi ve etik denetim eksiklikleri gibi riskler de barındırmaktadır.

Sonuç: Sonuç olarak, bu çalışma sağlık alanında LLM'lerin kullanımına yönelik kapsamlı bir haritalama sunarak, öne çıkan araştırma temalarını ve literatürdeki boşlukları ortaya koymaktadır.

Anahtar Kelimeler: Bibliyometrik analiz, ChatGPT, üretken yapay zeka, büyük dil modelleri, tıp

Corresponding Author: Hakan Öztürk ✉ h.ozturk3509@gmail.com **Received:** 03.07.2025 **Revision:** 28.11.2025 **Accepted:** 24.12.2025

Highlights

- The study identifies influential sources and emerging hotspots in healthcare-related Large Language Models research.
- WoS-based mapping of 2023-early 2025 publications characterizes the post-ChatGPT landscape of LLM research in healthcare.
- Output and collaborations cluster in major hubs (notably the USA and China) and high-capacity institutions, while Turkey remains peripheral in co-authorship networks.
- Core themes focus on AI/ChatGPT and performance; acceptance, risk, and equity/governance remain less mature, indicating key gaps and priorities for future work.

INTRODUCTION

Large Language Models (LLMs), in parallel with recent advances in natural language processing, offer substantial potential for health sciences and their applications (1). As healthcare increasingly depends on digital tools and data-driven decision-making, understanding how LLMs contribute to clinical, educational, and administrative settings becomes essential. The ability of these models to analyze and interpret complex data has the potential to transform the way healthcare is delivered, from diagnosis to treatment, from drug discovery to personalized medicine. Their superior performance in natural language processing tasks is particularly advantageous in analyzing text-based health data (1). LLMs can improve patient compliance by making medical instructions more understandable and reducing communication errors (2). By analyzing data from patient records, LLMs can improve adherence to medical prescriptions and minimize misunderstandings between healthcare professionals and patients. These models can also improve health literacy by rephrasing medical information in a language that non-experts can understand, helping patients make more informed decisions about their health (3). LLM applications in health sciences are not only limited to patient care and communication, but also play an important role in developing clinical decision support systems, predicting disease progression and accelerating medical research (4).

Since large language models, such as ChatGPT and other generative artificial intelligence (AI) models, are trained on massive amounts of health and medical data, they can develop specialized knowledge specific to different medical disciplines. In this way, they can make significant contributions to healthcare professionals in areas such as clinical decision support, early diagnosis of diseases and drug discovery. By analyzing the information in the medical literature, LLMs can specialize in different medical fields such as radiology, pathology and oncology and follow the latest developments in these fields (4,5). However, the potential benefits of LLMs in the healthcare field also come with some risks and limitations that need to be carefully considered. Although LLMs have human-like abilities to analyze and interpret medical information, they may not always produce accurate and reliable results (6). It should be kept in mind that the texts produced by these models may be incompatible with clinical and societal values and may contain misleading medical information (7). The use of LLMs, especially in clinical tasks, requires extensive expert training and rigorous validation processes. The use of models such as ChatGPT by the general public as a substitute for professional medical advice carries potential risks. Developers, healthcare professionals, ethics experts and regulatory authorities should act in collaboration (5). Transparency, accountability and ethical principles should be observed in the application of LLMs in health (8,9).

Bibliometric analysis is a method that aims to examine research trends, collaboration networks and academic impacts based on quantitative data of scientific publications. It maps knowledge production processes through criteria such as publication numbers, citation data, inter-author relationships and keyword analysis. This method helps to understand the body of knowledge in a particular field and identify prominent themes and gaps. Especially in rapidly changing disciplines such as health sciences, it provides researchers with an important tool for both systematically reviewing existing literature and providing strategic directions for future research (10,11).

Despite the rapidly increasing number of studies on LLM research in healthcare, comprehensive evaluations that systematically map the intellectual structure, thematic evolution, collaboration patterns, and emerging hotspots of the field remain limited. Existing publications primarily examine technical capabilities, ethical considerations, or specific application areas, but they do not offer an overarching, data-driven picture of how the field is evolving over time. In particular, research focusing on the post-ChatGPT era (2023 onwards), when the volume and diversity of publications noticeably increased, is still limited (12,13). Although

the field has expanded rapidly, the extent and characteristics of this acceleration have yet to be systematically documented. Therefore, this study aims to address this gap by conducting a comprehensive bibliometric analysis of LLM-related publications in the health sciences, identifying major research themes, collaboration networks, influential authors, and emerging trends.

MATERIAL and METHODS

Data Sources, Data Collection and Search Strategies

This research is a descriptive study in which studies published in the Web of Science (WoS) database between 2023 and 2025 on the use of LLM models in the field of health sciences are examined by bibliometric analysis method. The year 2023 was chosen as the starting point because it coincides with the period following the public release of widely accessible generative AI tools such as ChatGPT, when applications of LLMs in healthcare began to expand rapidly. The search query was constructed using key terms reflecting the scope of the study, including "Large Language Model*", "LLM*", "ChatGPT", "Generative AI", "Health", "Healthcare", "Medical", "Medical Practice", "Clinical Practice" and "Medicine". The search was conducted on the WoS database in a single session on 22.02.2025.

In the first screening, 1458 articles were found. The inclusion criteria were that they were published between 2023 and 2025, open access and full-text available in English. Exclusion criteria were articles without access to the full text, review articles and irrelevant studies based on manual evaluation of titles and abstracts. As a result, publications from 2025 that had been accepted or were in press but had not yet been indexed in WoS by 22.02.2025 could not be captured, which may lead to underestimation of the total number of publications for that year.

The flowchart created by taking into account the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines is presented in Figure 1. In this diagram, the boxes on the main vertical axis show the number of records retained at each PRISMA phase (identification, screening, eligibility and inclusion), while the boxes on the right-hand side indicate the reasons for exclusion and how many records were removed at each step. In the first search, 818 articles were accessed by selecting open access articles from the 1458 articles. Then, review articles (n=81) and enriched cited references (n=198) were removed and the years 2023-2024 and 2025 were selected. Thus, 539 articles were accessed. When the document type "article" was selected among these, 173 articles were identified.

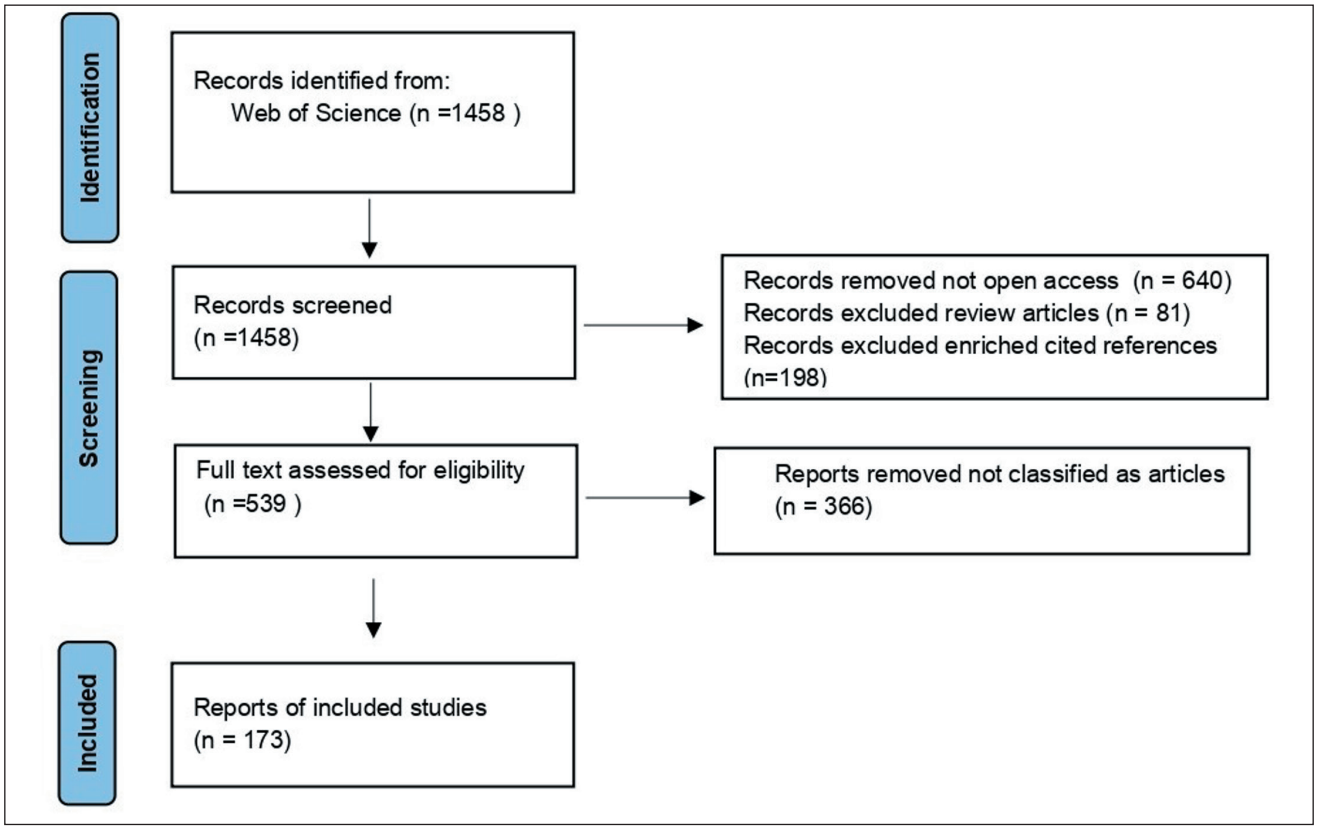

Figure 1: PRISMA flow diagram of the study selection process for LLM-related publications in healthcare

When these articles were analyzed according to Web of Science indexes, it was determined that 103 articles were published in SCI indexed journals, 6 articles in SSCI indexed journals and 64 articles in ESCI indexed journals.

The bibliometric analysis was conducted using the Biblioshiny web interface of the Bibliometrix R package. Analyses were conducted in R using the Bibliometrix package and its Biblioshiny graphical interface to perform both descriptive bibliometrics (e.g., annual scientific production, most productive authors, sources and countries) and science mapping (e.g., co-authorship, co-citation and keyword co-occurrence networks). Bibliometrix provides a comprehensive framework for importing bibliographic data from various academic databases, including WoS, and performing bibliometric analyses (11).

In this study, various bibliometric indicators were utilized, including document analysis to identify the most frequently used terms in the literature, keyword cloud visualization to highlight prominent topics, and keyword co-occurrence analysis to detect emerging research trends. Co-authorship (authors, institutions, countries) and keyword co-occurrence networks were generated using full counting, and the association strength normalization method was applied to construct the similarity matrices. For network visualization and community detection, the clustering procedure implemented in Bibliometrix/Biblioshiny (Louvain algorithm) was used to identify groups of closely related authors, countries and topics. To enhance interpretability and reduce visual noise, only items with at least three occurrences (for keywords) or at least two documents (for authors and countries) were included in the network maps, and isolated nodes were excluded. This approach allowed for a detailed examination of the evolution and significance of the topics covered in the academic literature.

RESULTS

As a result of the bibliometric analysis of 173 articles that met the inclusion criteria in the study, it was determined that these articles were published in 93 different sources, the number of authors working on this subject was 961, and the number of single-author articles was 13.

Most Relevant Sources

"Most Relevant Sources", which shows the journals with the highest number of publications and citations in the relevant academic field, is given in Table 1. In this study, the term "most relevant sources" refers to journals with the highest number of LLM-related articles within the analyzed WoS corpus. Table 1 shows the most important academic journals ranked according to the number of publications and the number of articles in these journals. Specifically, Table 1 presents the top 10 journals that published LLM-related articles in health sciences, ranked in descending order according to the number of articles identified in our dataset. The column "Sources" lists the names of the journals, while the column "Articles" indicates the number of publications from each journal. Accordingly, Journal of Medical Internet Research (JMIR) stands out as the journal with the highest number of articles published with 18 publications. Cureus Journal of Medical Science (14 publications) and JMIR Medical Education (13 publications) rank second and third, respectively. IEEE Access (5 publications), JMIR Mental Health (5 publications) and PLOS ONE (5 publications) also make notable contributions.

Most Relevant Affiliations

The institutions (affiliations) that produce the highest number of scientific publications in the relevant academic field are presented in Table 2. Table 2 presents universities and research institutions, along with the number of LLM-relat-

Table 1: Most Relevant Sources: Distribution of LLM-related Scientific Publications by Journal

Sources	Articles
Journal of Medical Internet Research	18
Cureus Journal of Medical Science	14
JMIR Medical Education	13
IEEE Access	5
JMIR Mental Health	5
Plos One	5
JMIR Medical Informatics	4
Proceedings of the ACM on Interactive Mobile Wearable and Ubiquitous Technologies-IMWUT	4
AI	3
JAMA Network Open	3

JMIR: Journal of Medical Internet Research, **IEEE:** Institute of Electrical and Electronics Engineers, **JAMA:** Journal of the American Medical Association

Table 2: Most Relevant Institutions

Affiliation	Articles
Harvard University	16
University of California System	15
Feinberg School of Medicine	12
Northwestern University	12
University of Ulsan	12
Vanderbilt University	12
Sichuan University	10
Chongqing Medical University	9
Harvard University Medical Affiliates	9
Harvard Medical School	8

ed publications produced by each within the analyzed WoS dataset. More specifically, it shows the top 10 affiliations, ordered in descending number of articles. The column "Affiliation" displays the institution name as indexed in WoS, while the column "Articles" reports the number of publications from our final corpus associated with that institution. Table 2 reveals which academic institutions conduct more research in this field. Harvard University stands out as the academic institution that produces the most publications with 16 publications. The University of California System ranks second with 15 publications. The leadership of these two institutions reflects the leading role of the US in scientific productivity. Other leading universities, Feinberg School of Medicine, Northwestern University, University of Ulsan and Vanderbilt University, have a remarkable scientific production with 12 publications. These institutions are particularly active in the fields of medicine, health informatics and artificial intelligence applications. China-based universities such as Sichuan University (10 publications) and Chongqing Medical University (9 publications) demonstrate the strength of academic research in Asia. This finding reflects China's growing scientific activity in the fields of medical informatics and artificial intelligence. Harvard University Medical Affiliates (9 publications) and Harvard Medical School (8 publications) show the scope of Harvard University's research in medicine and health informatics. These results are important to understand which institutions are driving scientific research and to identify academic centers for future collaboration.

Most Relevant Words

Findings on the most relevant words that are prominent in academic publications are given in Table 3. In Table 3, the "Words" column lists the author keywords, while the "Occurrences" column shows how many times each keyword appears in the dataset. Table 3 provides an understanding of the trends in a particular field by revealing which topics are emphasized more in academic studies. "Artificial Intelligence" stands out as the most frequently used keyword with 16 repetitions. "ChatGPT" ranks second with 10 occurrences. This finding shows that AI and ChatGPT are increasingly being examined in academic research and the use of these technologies in various fields is being investigated. The word "Education" was used 8 times and the word "Performance" was used 7 times. This shows the growing academic interest in how AI is being used, especially in the context of education.

Table 3: Most Relevant Keywords

Words	Occurrences
artificial-intelligence	16
chatgpt	10
education	8
performance	7
challenges	5
accuracy	4
AI	4
depression	4
health	4
impact	4

AI: Artificial intelligence

Tree Map

The tree map visualizing the most frequently used keywords in scientific studies and their importance in the academic literature is given in Figure 2. The size of each box reflects the number of repetitions of the relevant word and shows which themes are at the forefront of scientific research. Each rectangle represents one author keyword, and the different colors are used only to visually distinguish the keywords, while the numbers and percentages inside each box indicate the absolute frequency and relative share of that keyword in the dataset. According to Figure 2, "Artificial Intelligence" is the keyword that ranks first with 82 repetitions (19%). "ChatGPT" ranked second with 63 occurrences (14%). These findings show that artificial intelligence and large language models are becoming an increasingly central topic in scientific research. "Large Language Models" has an important place with 48 occurrences (11%). The terms "LLM" and "Large Language Model" were used 33 times in total. The term "Generative AI" was used 11 times. The term "medical education" was used 32 times (7%). "Chatbots" is included in the literature as one of the applications of large language models with 8 repetitions.

Thematic Map

The thematic map presented in Figure 3 visualizes the most common themes and their importance levels in the scientific literature through bibliometric analysis. Figure 3 is divided into four sections where keywords are classified in terms of their density and centrality. On the horizontal axis, the "Relevance degree (Centrality)" indicates how strongly a theme is connected to other themes in the overall network, while on the vertical axis the "Development degree (Density)" reflects the internal cohesion and maturity of each theme. The size of the circles represents the frequency of the keywords in the dataset, and the different colors indicate distinct thematic clusters. These are Motor Themes (Upper Right Quadrant), Core Themes (Lower Right Quadrant), Niche Themes (Upper Left Quadrant) and Emerging or Weakening Themes (Lower Left Quadrant).

In the motor themes region, the most powerful and guiding themes in the scientific field with high centrality and

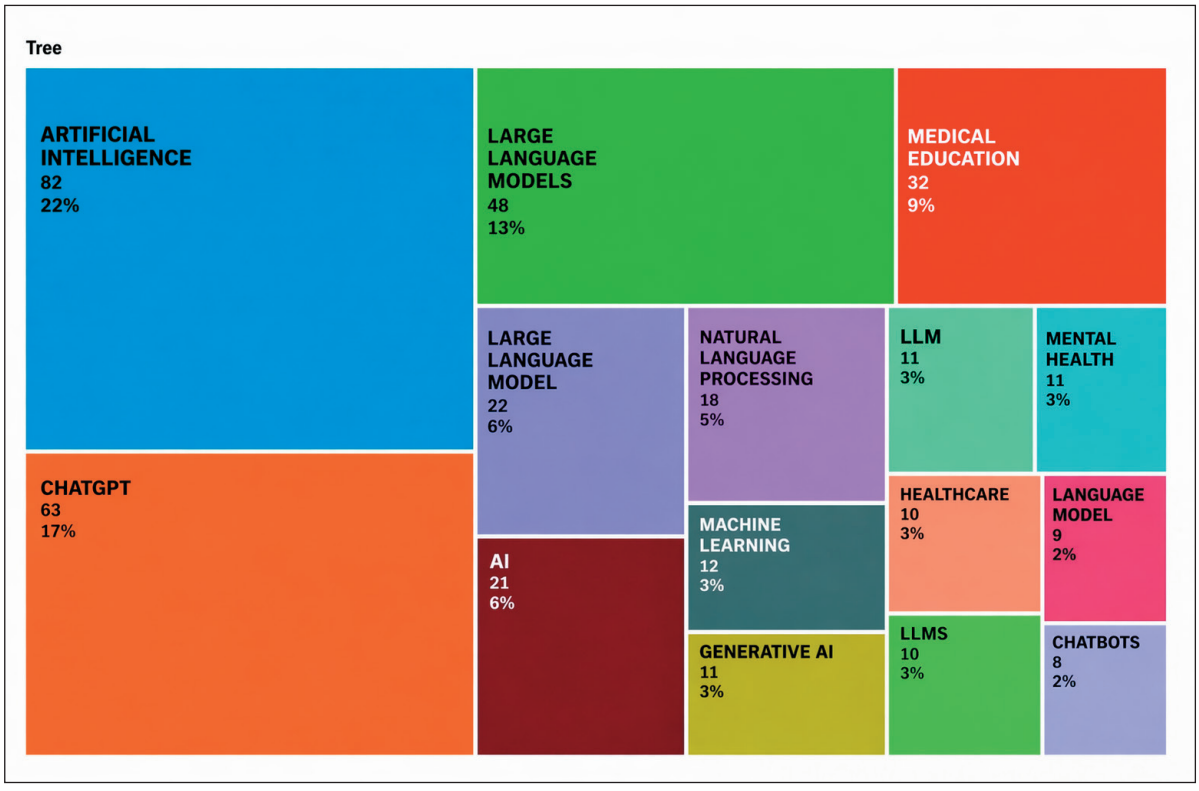

Figure 2: Treemap of the most frequently used author keywords in LLM-related healthcare publications

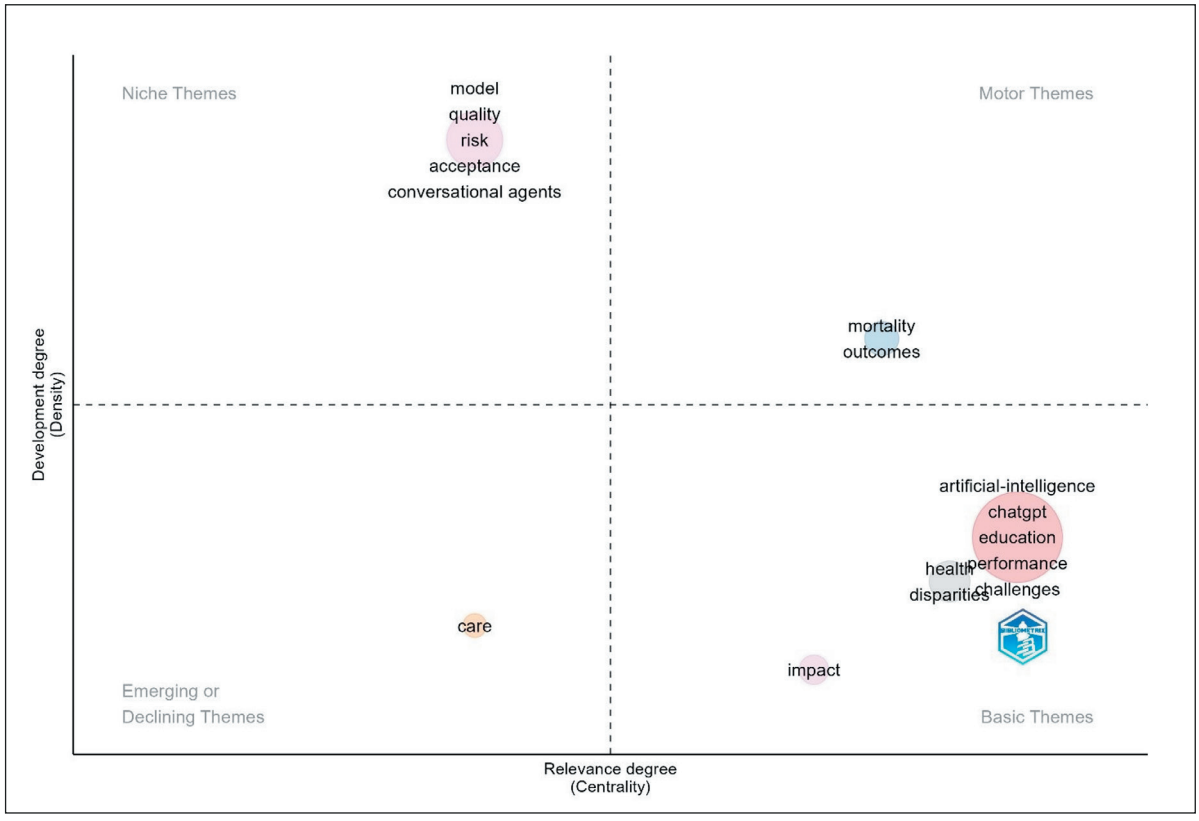

Figure 3: Thematic Map

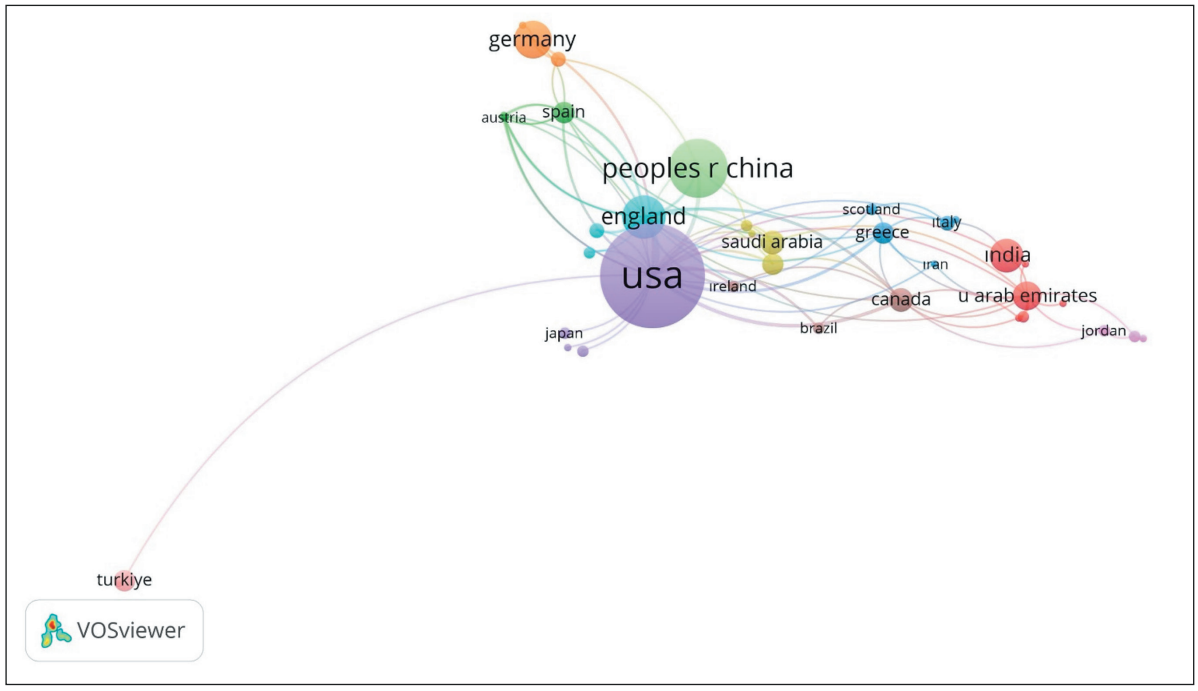

Figure 4: Academic Collaboration Network

sophistication in the literature are located. "mortality" and "outcomes" are located in this region, but do not form a large cluster. This shows that mortality and clinical outcomes remain an important area in health and medical research, but topics such as AI and ChatGPT are becoming more central.

Core Themes include topics that have been widely studied and are central to the literature but are not yet fully developed. "artificial intelligence", "chatgpt", "education", "performance", "health disparities", "challenges" and "impact" are included in this section. This finding shows that artificial intelligence and chatgpt have become a core academic topic, especially in the context of education and health. Issues such as health inequalities and performance are indicative of the growing interest in how AI-enabled systems are being used in health.

Niche Themes include topics that are more specific and intensively studied but not central to the literature. "model", "quality", "risk", "acceptance" and "conversational agents" fall into this category. This shows that academic work on conversational agents (e.g. chatGPT) and the reliability, risks and acceptance of language models is intensifying. However, these topics have not yet become a large-scale central theme in the literature.

Emerging or weakening themes include the concept of "care". This suggests that work on care services of AI and ChatGPT is still developing or has not made a strong impact in the literature.

Academic Collaboration Network

In order to analyze scientific collaborations between countries, a bibliometric map derived from co-authorship data is presented in Figure 4. This map was created using VOSviewer software, collaborative relationships between countries are represented by nodes and edges. Node sizes vary depending on the number of co-authorships of the respective country, with larger nodes indicating more collaboration. In this visualization, the thickness of the connecting lines ("edges") reflects the strength of collaboration between two countries (number of co-authored documents), and different colors indicate distinct clusters of closely collaborating countries. Figure 4 shows that the node with the largest and most densely connected network is USA. Accordingly, it can be said that the USA has a central position in international scientific collaborations and collaborates with a large number of countries. Countries such as the UK, China and Germany also stand out as important nodes in the network. These countries have a wide network of cooperation and are connected to a large number of countries. In Europe, there are strong links between the UK, Germany and Spain. Countries such as India and the United Arab Emirates have developed closer ties with specific regions. Turkey is positioned far from the network and has relatively few connections in global collaborations. As noted above, the different colors correspond to clusters of countries that are more closely connected to each other than to countries

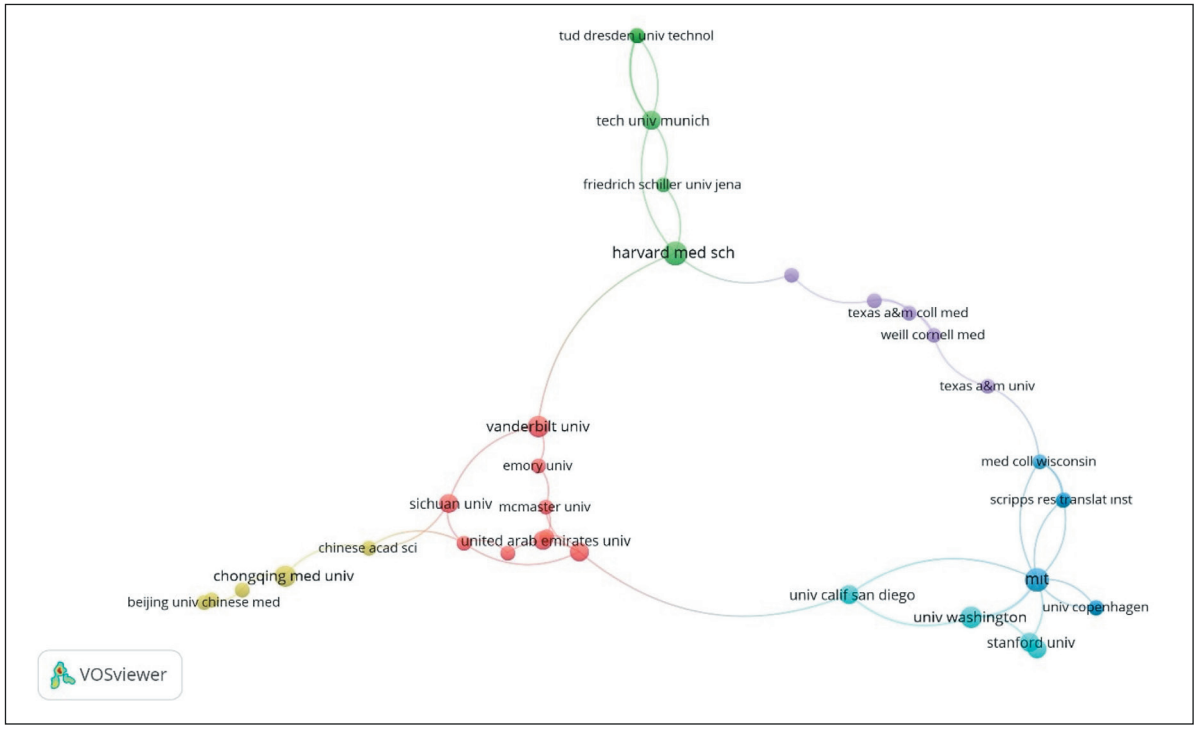

Figure 5: Cooperation Network between Academic Institutions

in other clusters. The blue-green network, which includes the US and the UK, shows strong academic collaborations between developed countries, while the red-pink network, which includes China and India, presents a more regional model of collaboration. This analysis reveals that the US has a central position in scientific research, while the UK, Germany and China stand out as important cooperation actors.

Co-authorship Analysis

As a result of the co-authorship analysis conducted in the study, Figure 5 presents a map of the collaborations between universities and research centers. Each circle (node) represents a university or research institution, the size of the node is proportional to the number of co-authored documents, the connecting lines (edges) indicate co-authorship links, and different colors denote clusters of institutions that collaborate more frequently with each other. Accordingly, one of the most important nodes is Harvard Medical School, shown in green, which is connected to various academic institutions from different geographies. This shows that Harvard has a wide network of academic collaborations at the international level. In particular, it is directly linked to three major universities in Germany (TU Dresden, Technical University of Munich and Friedrich Schiller University Jena), indicating that it also attaches importance to collaborations in Europe. Leading US universities such as the Massachusetts Institute of Technology (MIT), the University of Washington and Stanford University are notable for their strong collaboration. When regional clusters and academic collaboration are examined, China-based institutions (Sichuan University, Chongqing Medical University, Beijing University of Chinese Medicine and Chinese Academy of Sciences), which can also be called the Asian cluster and shown in yellow in the figure, form a strong collaboration network within themselves. This may reflect the influence of regional science policies and local research funding. The Middle East and Canada Collaboration is also shown in red in the figure, with a notable academic partnership between the United Arab Emirates University, McMaster University and Emory University. This cluster represents the academic ties between Canada, the US and the Middle East. Another noteworthy element in the figure is the relatively low number of direct links between China-based academic institutions and major research universities in the US. This suggests that regional trends in scientific collaborations are strong and that some academic barriers may exist globally.

DISCUSSION

This study reveals that academic productivity towards Large Language Models in healthcare has increased significantly in recent years and research trends are concentrated around specific themes. By focusing on WoS-indexed publications between 2023 and early 2025, our analysis quantitatively characterizes the “post-ChatGPT” period and shows how scientific output and collaboration are distributed across

countries, institutions and research topics. The results of the bibliometric analysis show that the United States, China and some European countries are leading the way in this field and that scientific collaboration networks are shaped around these countries. The leadership of institutions such as Harvard University, University of California System and Northwestern University at both the publication and collaboration level reveals that the central actors of AI research in healthcare are largely institutions with high research capacity (14,15). Compared with previous narrative and scoping reviews of LLMs in healthcare, the present study adds a data-driven mapping of where this research is produced and which institutions occupy structurally central positions in the global collaboration network.

The fact that the most frequently encountered keywords include “Artificial Intelligence”, ‘ChatGPT’, “Medical Education” and “Performance” shows that LLM technologies are widely used in health systems not only in decision support processes but also in areas such as health education, communication and service quality (1,16). Our treemap results further quantify this pattern by showing that “artificial intelligence” and “ChatGPT” together account for a substantial proportion of all keywords, confirming that LLMs have rapidly become a focal topic in health-related AI research. Tree map and thematic map analyses show that AI-based approaches also touch social dimensions such as “health disparities” and “challenges”. These findings indicate that LLM applications should be evaluated not only in technological but also in socio-ethical contexts (8,17). In this respect, our thematic structure, in which terms related to equity and challenges appear alongside performance-oriented keywords, empirically supports the growing emphasis on ethical and societal implications reported in recent literature. Similarly, a recent national bibliometric study from Turkey has highlighted ethical and governance-related themes in AI applications in healthcare, underscoring the importance of these issues in the local context as well (18).

According to the Thematic Map analysis, concepts such as “artificial intelligence” and “ChatGPT” appear as core themes, while terms such as ‘acceptance’, “conversational agents” and “risk” are more niche or emerging areas. This suggests that dimensions such as user perception, trust, and ethical awareness are still immature research areas in the adoption process of LLM technologies (19,20). Our finding that these topics cluster as niche or emerging themes indicates that, despite intense conceptual and opinion-based discussion in the literature, empirical and large-scale studies on acceptance, risk perception and real-world deployment of conversational agents in healthcare remain relatively limited. Many studies emphasize that the use of LLM, especially in clinical settings, should be evaluated within the framework of security, accuracy and privacy (21,22), and our keyword and thematic analyses complement these concerns by demonstrating that such issues are beginning to gain visibility but have not yet matured into firmly established core themes.

In the analysis of cooperation between countries, it is seen that countries such as the USA, the UK, China and Germany are pioneers in scientific productivity and have extensive cooperation networks. This shows that LLM-based health research is largely concentrated in specific scientific centers (10). It is important for developing countries such as Turkey to develop more international collaboration in order to utilize these technologies effectively and reliably in the local context. In our co-authorship map, Turkey appears as a small, weakly connected node located at the periphery of the global network, indicating limited participation in large, multicenter LLM projects. This peripheral position may reduce the visibility and impact of national research outputs and slow down the transfer of methodological innovations and best practices into the local health system. Strengthening collaboration with leading centers in North America, Europe and Asia would therefore be strategically important for capacity building, access to high-quality datasets and the development of context-appropriate LLM applications for the Turkish healthcare system.

The results of this study are in line with the findings in the literature on the role of LLMs in healthcare. For example, Liu et al. (1) state that LLMs are effective in diagnostic support, medical summarization, clinical question answering and patient interaction; however, more research is needed on issues such as ethics, transparency and model interpretability. Similarly, Bi et al. (14) state that LLMs can contribute to multimodal data integration in healthcare, but data standardization and reliability issues are still important challenges to be solved. Our mapping refines these observations by showing that performance-related terms (e.g. “performance”, “accuracy”, “outcomes”) are already central in the keyword structure, whereas issues such as data quality, standardization and explainability are present but less prominent, suggesting a gap between the technical potential of LLMs and the maturity of research on their safe and robust implementation.

It is also emphasized that generative models such as ChatGPT offer potential in curriculum development, simulation and assessment processes in medical education, but cannot replace direct human interaction with faculty members (19,20). It is suggested that such technologies should be considered as supportive and complementary tools in educational practices. Consistent with this, our results show that “medical education” and related terms form an important thematic area, indicating that educational applications are one of the most actively explored domains for LLMs in health sciences; however, other application areas such as long-term outcome evaluation and health service delivery

appear less developed in the thematic map, highlighting opportunities for future empirical research.

However, it is also stated that LLM-based systems have serious limitations such as "hallucination" (generating false information), misjudging the context, creating user bias and unethical guidance (8,22). It is clear that such systems should only be used in a supervised and transparent environment, especially in clinical applications. As a matter of fact, institutions such as the World Health Organization (WHO) also emphasize that artificial intelligence applications in health should be supervised within an ethical framework (23). The presence of keywords such as "challenges", "health disparities" and "risk" in our analysis is consistent with these concerns and suggests that the research community has started to incorporate ethical and governance issues into the LLM agenda; nevertheless, their positioning as non-motor themes indicates that systematic, outcome-oriented evaluations of ethical frameworks and governance models are still relatively scarce.

Conclusion

This study provides a comprehensive mapping of the use of LLMs in healthcare, revealing the direction of academic interest, strong themes and existing gaps. Our findings highlight three main contributions: (i) a detailed description of the global and institutional landscape of LLM-related health research in the immediate post-ChatGPT period, (ii) an identification of core and emerging themes that structure the current literature, and (iii) an evidence-based demonstration of collaboration asymmetries, particularly the peripheral position of some developing countries such as Turkey in international co-authorship networks. Multifaceted, interdisciplinary and ethics-based research approaches are needed to integrate LLM technologies into health systems. Future studies should therefore prioritize rigorous, data-driven evaluations of clinical effectiveness, safety and equity impacts of LLM applications, as well as strategies to broaden international collaboration and reduce geographic and thematic imbalances identified in this bibliometric analysis.

Limitations

This study has several limitations that should be considered when interpreting the findings. First, the analysis includes only Web of Science-indexed, open-access, English-language articles, which may underrepresent studies from other databases, subscription journals or non-English sources. Second, the time window is restricted to 2023-22 February 2025, so publication counts for 2025 are provisional and later-indexed articles are not captured. Third, the search strategy relied on a predefined set of LLM-related keywords (e.g. "Large Language Model", "LLM", "ChatGPT", "Generative "), meaning that studies using different terminology may have been missed.

Author Contributions

Study conception and design: **Hakan Öztürk, Elvan Hayat;** data collection: **Hakan Öztürk, Elvan Hayat;** analysis and interpretation of results: **Hakan Öztürk, Elvan Hayat;** draft manuscript preparation: **Hakan Öztürk, Elvan Hayat.** The authors reviewed the results and approved the final version of the article.

Conflicts of Interest

The authors declare that there is no conflict of interest to disclose.

Financial Support

The authors declare that the study received no funding.

Ethical Approval

Ethical approval was not required for this study as it is a bibliometric analysis based on publicly available data.

Review Process

Extremely and externally peer-reviewed.

REFERENCES

1. Liu J, Yang M, Yu Y, Huang X, Zhang Q. Large language models in bioinformatics: applications and perspectives. arXiv. 2024;2401.04155.
2. Adedeji A, Sanni M, Ayodele E, Okoro M, Bello A. The multicultural medical assistant: can LLMs improve medical ASR errors across borders? arXiv. 2025;2501.15310. https://doi.org/10.48550/arXiv.2501.15310
3. Wen-Ching Wu MSN. Effects of a Steno Diabetes Dialogue Card-based intervention on self-perceived health, health literacy, and glycemic control in older adults with type 2 diabetes mellitus living in the community. Hu Li Za Zhi. 2025;72(1):64-75. https://doi.org/10.6224/JN.202502_72(1).09
4. Karabacak M, Margetis K. Embracing large language models for medical applications: opportunities and challenges. Cureus. 2023;15(12):e39305. https://doi.org/10.7759/cureus.39305
5. Vrdoljak J, Boban Z, Vilović M, Kraljević T, Petrović L. A review of large language models in medical education, clinical decision support, and healthcare administration. Healthc (Basel). 2025;13(6):603. https://doi.org/10.3390/healthcare13060603
6. Alber DA, Yang Z, Alyakin A, Brown P, Chen L. Medical large language models are vulnerable to data-poisoning attacks. Nat Med. 2025;31:1-9. https://doi.org/10.1038/s41591-024-03445-1
7. Tangsrivimol JA, Darzidehkalani E, Virk HU, Smith J, Lee V. Benefits, limits, and risks of ChatGPT in medicine. Front Artif Intell. 2025;8:1518049. https://doi.org/10.3389/frai.2025.1518049
8. Clusmann J, Kolbinger FR, Muti HS, O'Connor C, Peters G. The future landscape of large language models in medicine. Commun Med. 2023;3(1):141. https://doi.org/10.1038/s43856-023-00370-1
9. Kumar K, Ashraf T, Thawakar O, Anwer RM, Singh P. LLM post-training: a deep dive into reasoning large language models. arXiv. 2025;2502.21321. https://doi.org/10.48550/arXiv.2502.21321

10. Donthu N, Kumar S, Mukherjee D, Pandey N, Lim WM. How to conduct a bibliometric analysis: an overview and guidelines. J Bus Res. 2021;133:285-296. https://doi.org/10.1016/j.jbusres.2021.04.070
11. Aria M, Cuccurullo C. Bibliometrix: an R-tool for comprehensive science mapping analysis. J Informetr. 2017;11(4):959-75. https://doi.org/10.1016/j.joi.2017.08.007
12. Barrington NM, Gupta N, Musmar B, Doyle D, Panico N, Godbole N, et al. A bibliometric analysis of the rise of ChatGPT in medical research. Med Sci (Basel). 2023;11(3):61. https://doi.org/10.3390/medsci11030061
13. Koo M. ChatGPT research: a bibliometric analysis based on the Web of Science from 2023 to June 2024. Knowledge. 2025;5(1):4. https://doi.org/10.3390/knowledge5010004
14. Bi Z, Dip SA, Hajialigol D, Kommu S, Liu H, Meng L, Patel R. AI for biomedicine in the era of large language models. arXiv. 2024;2403.15673. https://doi.org/10.48550/arXiv.2403.15673
15. Gencer G, Gencer K. Large language models in healthcare: a bibliometric analysis and examination of research trends. J Multidiscip Healthc. 2025;18:223-238. https://doi.org/10.2147/JMDH.S502351
16. Cascella M, Semeraro F, Montomoli J, Rossi A, Bianchi F. The breakthrough of large language models release for medical applications: 1-year timeline and perspectives. J Med Syst. 2024;48(1):22. https://doi.org/10.1007/s10916-024-02045-3
17. Lawrence HR, Schneider RA, Rubin SB, Davis K, Williams L. The opportunities and risks of large language models in mental health. JMIR Ment Health. 2024;11(1):e59479. https://doi.org/10.2196/59479
18. Çelik Ö, Kaya E. Artificial intelligence and ethics in healthcare: a bibliometric analysis. Süleyman Demirel University Visionary Journal. 2024;15(43):1046-1062. https://doi.org/10.21076/vizyoner.1455659
19. Abd-Alrazaq A, AlSaad R, Alhuwail D, Househ M, Shah Z. Large language models in medical education: opportunities, challenges, and future directions. JMIR Med Educ. 2023;9(1):e48291. https://doi.org/10.2196/48291
20. Khan AA, Akbar MA, Fahmideh M, Rahman N, Li C. AI ethics: an empirical study on the views of practitioners and lawmakers. IEEE Trans Comput Soc Syst. 2023;10(6):2971-2984. https://doi.org/10.1109/TCSS.2023.3251729
21. Yu D, Bondi M, Hyland K. Can GPT-4 learn to analyse moves in research article abstracts? Appl Linguist. 2024;amae071. https://doi.org/10.1093/applin/amae071
22. Zhou Z, Li J, Zhang Z, Chen F, Wu X. Examining how the large language models impact the conceptual design with human designers: a comparative case study. Int J Hum-Comput Interact. 2024;1-17. https://doi.org/10.1080/10447318.2024.2370635
23. World Health Organization. Ethics and governance of artificial intelligence for health: guidance on large multi-modal models. Geneva: WHO; 2024. ISBN: 978-92-4-008475-9