frontiers | Frontiers in Medicine

TYPE Systematic Review
PUBLISHED 01 July 2025
DOI 10.3389/fmed.2025.1456286

OPEN ACCESS

EDITED BY
Md. Mohaimenul Islam,
The Ohio State University, United States

REVIEWED BY
Francesco Monaco,
Azienda Sanitaria Locale Salerno, Italy
TaChen Chen,
Nihon Pharmaceutical University, Japan

*CORRESPONDENCE
Chunyan Han
✉ 15552869804@163.com

†These authors have contributed equally to this work and share first authorship

RECEIVED 28 June 2024
ACCEPTED 13 May 2025
PUBLISHED 01 July 2025

CITATION
Yao L, Liu Y, Wang T, Han C, Li Q, Li Q, You X, Ren T and Wang Y (2025) Global trends of big data analytics in health research: a bibliometric study.
Front. Med. 12:1456286.
doi: 10.3389/fmed.2025.1456286

Global trends of big data analytics in health research: a bibliometric study

Li Yao[1,2†], Yan Liu[3†], Tingrui Wang[3], Chunyan Han[4*], Qiaoxing Li[2], Qinqin Li[3], Xiaoli You[1], Tingting Ren[5] and Yinhua Wang[6]

[1]Department of Respiratory and Critical Care Medicine, The Affiliated Hospital of Guizhou Medical University, Guiyang, China, [2]School of Management & Collaborative Innovation Laboratory of Digital Transformation and Governance, Guizhou University, Guiyang, China, [3]School of Nursing, Guizhou Medical University, Guiyang, China, [4]Emergency Critical Care Unit, Qingdao Municipal Hospital Group, Qingdao, China, [5]Department of Hepatobiliary Surgery, The Affiliated Hospital of Guizhou Medical University, Guiyang, China, [6]Department of Nursing, The Affiliated Hospital of Guizhou Medical University, Guiyang, China

Background: The field of "Big Health," which encompasses the integration of big data in healthcare, has seen rapid development in recent years. As big data technologies continue to transform healthcare, understanding emerging trends and key advancements within the field is essential.

Methods: We retrieved and filtered articles and reviews related to big data analytics in health research from the Web of Science Core Collection, including SCI Expanded and SSCI, covering the period from 2009 to 2024. Bibliometric and co-citation analyses were conducted using VOSviewer and CiteSpace.

Results: A total of 13,609 papers were analyzed, including 10,702 original research and 2,907 reviews. Co-occurrence word analysis identified six key research areas: (1) the application of big data analytics in health decision-making; (2) challenges in the technological management of health and medical big data; (3) integration of machine learning with health monitoring; (4) privacy and ethical issues in health and medical big data; (5) data integration in precision medicine; and (6) the use of big data in disease management and risk assessment. The co-word burst analysis results indicate that topics such as personalized medicine, decision support, and data protection experienced significant growth between 2015 and 2020. With the advancement of big data technologies, research hotspots have gradually expanded from basic data analysis to more complex application areas, such as the digital transformation of healthcare, digital health strategies, and smart health cities.

Conclusion: This study highlights the growing impact of big data analytics in healthcare, emphasizing its role in decision-making, disease management, and precision medicine. As digital transformation in healthcare advances, addressing challenges in data integration, privacy, and machine learning integration will be crucial for maximizing the potential of big data technologies in improving health outcomes.

KEYWORDS

big data, health, bibliometric study, VOSviewer, CiteSpace

1 Introduction

Big data involves the comprehensive analysis and processing of all available data, avoiding the simplifications inherent in random sampling surveys. Big data is traditionally defined by five characteristics: volume, velocity, variety, value, and veracity (1). Advancements in technology and deeper applications have broadened the characteristics of big data to include variability,

visualization, verifiability, value density, and viability (2). These characteristics not only highlight the challenges associated with volume and velocity in big data but also emphasize the potential to extract accurate, reliable, and valuable information from complex and variable data sets (3). As information technology continues to advance rapidly, big data has permeated various aspects of life, establishing an increasingly close connection with health. The emergence of the big data era presents an opportunity to manage dynamic health conditions, address health issues promptly, and develop personalized medical strategies (4). Clearly, health and medical big data have become critical areas of research.

Health and medical big data encompass all data related to medical care and health outcomes generated throughout the medical process (5). "This includes electronic health records, medical monitoring records, biometric data, public health information, and health insurance data. In healthcare and medical fields, big data can be extensively utilized for clinical decision support, pharmaceutical development, disease monitoring, and health management. This utilization involves various big data analytics techniques, including data structuring, image analysis, and intelligent detection (6). Consequently, the technologies associated with the application of big data are crucial to advancing the healthcare industry.

Recent studies have extensively explored the transformative role of big data analytics in medical practice. For instance, Lorenzo et al. (7) emphasized the transformative potential of big data in predicting oncology patient outcomes and enhancing personalized treatment. Another pivotal study by Dong et al. (8) demonstrated how big data facilitates real-time epidemic tracking, significantly aiding rapid and effective public health responses. Furthermore, Kindle et al. (9) explored the incorporation of big data analytics into clinical decision-making systems, discovering that data-based models significantly enhanced diagnostic precision and the efficiency of patient care. These studies highlight the essential role of advanced analytics in improving medical services and outcomes. Despite advancements, further detailed analyses and updates on the global implementation of big data and its long-term impact across various medical fields are still needed. Our research aims to address this issue through bibliometric methods.

Bibliometric research, which analyzes the characteristics of literature, serves as a technique for examining the distribution structure, quantitative relationships, and evolutionary patterns of relevant information within publications. This approach is used to assess research output and trends across various fields (10). VOSviewer, a Java-based software, enables the construction and visualization of bibliometric networks, such as citation coupling, co-citation analysis, author co-citation, and co-occurrence word analysis based on scientific publications (11). Bibliometric analysis using VOSviewer has been applied in various medical fields, including surgery (12), oncology (13), and nutrition (14), to gain deeper insights. Several bibliometric studies focusing on big data have been published, addressing topics such as infectious diseases (15), HIV (16), and critical care (17). Two studies have focused on bibliometric research in the healthcare industry using big data, analyzing articles published before 2016 (6, 18). Big data research in healthcare has garnered significant attention both domestically and internationally, prompting more researchers to use big data analytics tools to address medical issues. Since 2016, there has been a growing number of studies on the application of big data technologies in healthcare, which require further exploration through bibliometric analysis (19). In addition, CiteSpace is used to detect the knowledge structure, the evolution of research hotspots, and the burst trends of citations in the literature. This study aims to employ VOSviewer and CiteSpace for an updated and comprehensive analysis of publications on health big data analytics.

2 Materials and methods

2.1 Study design

A bibliometric analysis was conducted using VOSViewer version 1.6.19 and CiteSpace 6.2.R4 to investigate research on big data analytics in healthcare. The bibliometric methodology involved d five stages: study design, data collection, data analysis, data visualization, and interpretation (20).

2.2 Search strategy

Two investigators independently conducted a literature search. We searched the Web of Science (WoS) Core Collection, including SCI Expanded and SSCI, for studies published between 2009 and November 28, 2024. WoS was selected because of its extensive use in bibliometric studies and its superior coverage of high-impact journals (21). The search strategy employed was "TS = Topic." The search formulas used were TS = (big data) and TS = (health OR healthcare OR clinical OR medical OR medicine OR medical care).

2.3 Screening strategy

We excluded non-English articles, duplicate literature, letters, meeting abstracts, news items, editorials, comments, and retracted publications. There were no restrictions on publication date. Figure 1 illustrates the process and results of literature screening. A total of 13,609 documents were included in the analytic sample.

2.4 Data extraction

The extracted literature information for visualization and bibliometric analysis includes the publication year, journal title, authorship, WoS category, manuscript type, publication country/region, publication organization, total citations, and H5-index.

2.5 Statistical analysis

The exported bibliographic file was first imported into Occurrence 14.9 software for deduplication, data cleaning, and synonym consolidation. Subsequently, information such as publication date, authors, institutions, journals, and keywords was extracted. Bibliometric methods utilize mathematics, statistics, and philology to quantitatively analyze elements such as journal titles, publication years, countries/regions, organizations, authorship, citation counts, and H5-index. This study employed VOSviewer (version 1.6.19) to map keyword co-occurrence, citations, publications, bibliographic coupling in countries and institutions, as well as thematic and trend topic networks. CiteSpace 6.2.R4 was also used to detect bursts in keywords and references.

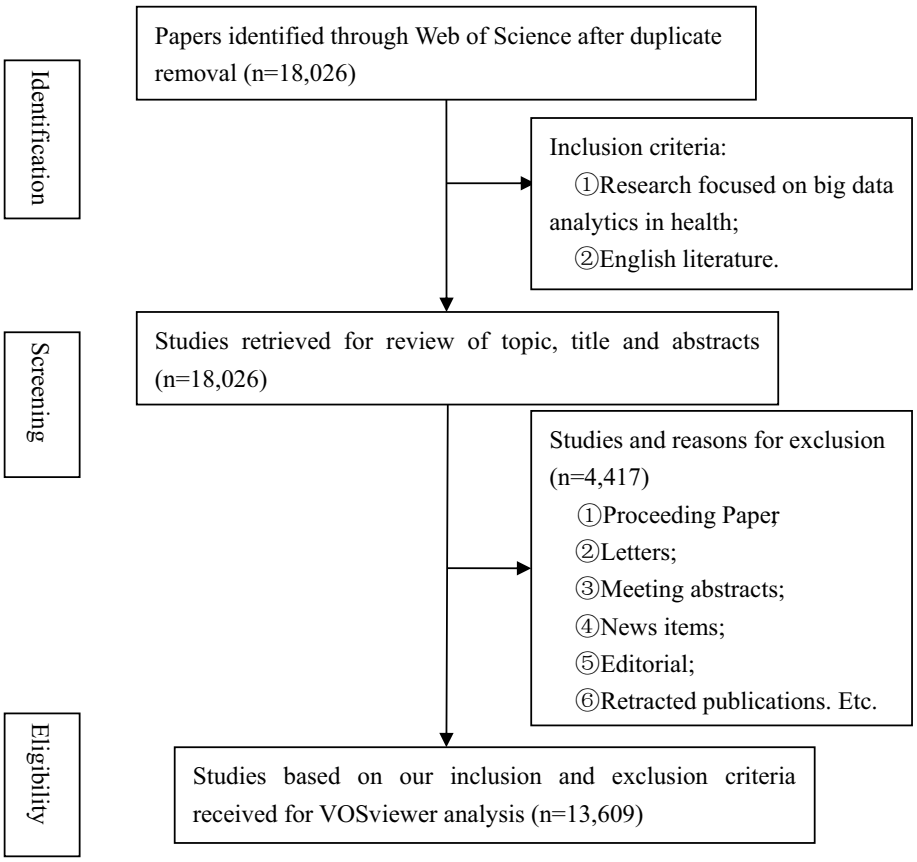

FIGURE 1
Flowchart of the preferred reporting items for VOSviewer analysis.

3 Results

3.1 Number of publications and trend analysis

Among the 13,609 articles included, 10,702 (78.6%) were original articles, and 2,907 (21.4%) were reviews. Scholars have published articles on the application of big data technology in the medical industry since 2009. Since 2014, literature on big data applications in this field has shown annual growth. The number of papers published in 2022 reached its peak at 2,286, which is 36.3 times greater than that published in 2013 and earlier. Polynomial fitting yielded an R^2 value of 0.893, indicating a significant correlation between the year of publication and the annual number of publications. This trend suggests that literature in this field is likely continue to expand. The application of big data analytics in the medical field remains a focal point of research. Figure 2 presents the contents.

3.2 Analysis of countries/regions

The literature in this field originates from 149 countries/regions. The USA and China have each published more than 3,000 articles, and 97 countries/regions have published more than five articles. Table 1 summarizes the top 10 countries/regions. The most prolific country was the USA, with 4,053 publications on medical big data analytics, which also had the highest total citations (159,464). Following China with 3,184 publications, England ranked third with 1,341 publications and the highest average citation rate of 40.06. The collaboration between countries and regions was analyzed using VOSviewer software (Figure 3). The lines between nodes indicate a cooperative relationship; the thicker the lines, the closer the relationship, denoting a stronger total link strength (TLS). The countries with the highest TLS were the USA (3,957), England (3,059), China (2,065), Germany (1,852), and Italy (1,705).

3.3 Analysis of institutions

Of the 13,959 institutions involved in research on big data analytics in medicine, 1,436 had published more than 5 papers. Table 2 presents the publications of the top 10 institutions over the last 15 years. Harvard Medical School (211 articles), the Chinese Academy of Sciences (209 articles), and Stanford University (169 articles) were the top three institutions in terms of published articles. The interinstitutional cooperation network map for institutions with 15 or more publications was generated using VOSviewer software (Figure 4). The leading institutions by TLS were Harvard Medical School (755), the University of Toronto (602), and the University of California, San Francisco (499).

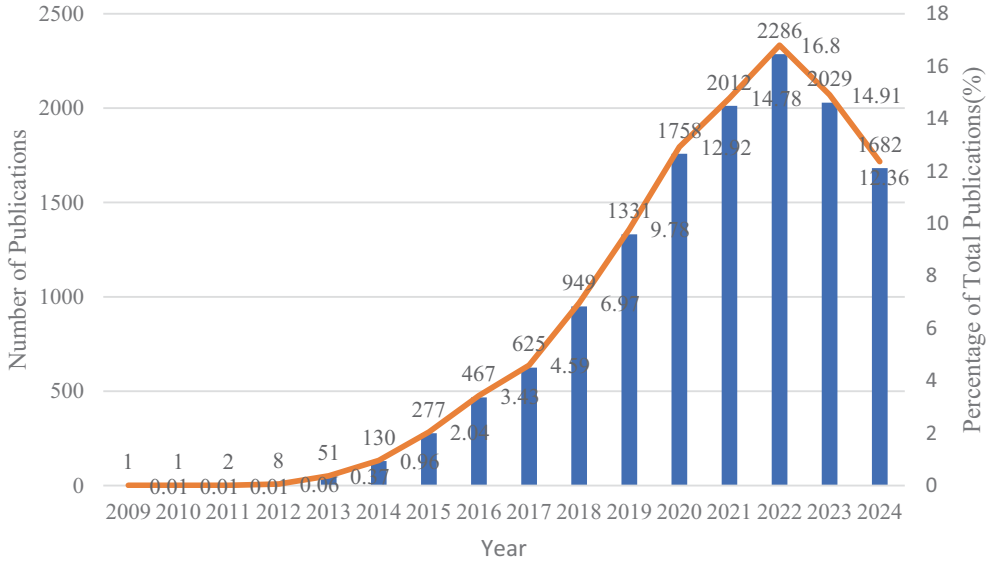

FIGURE 2
The annual publications on big data analytics in health from 2009 to 2024.

TABLE 1 The main countries/regions, and institutions contributing to publications on big data analytics in health.

Rank	Country/Region	Counts	Total citation	Average citation	Total link strength
1	USA	4,053	159,464	39.34	3,957
2	China	3,184	82,042	25.77	2,065
3	England	1,341	53,716	40.06	3,059
4	India	1,095	26,317	24.03	1,350
5	South Korea	827	19,231	23.25	784
6	Germany	770	21,436	27.84	1,852
7	Italy	759	20,870	27.50	1,705
8	Australia	753	27,097	35.99	1,637
9	Canada	735	27,569	37.51	1,374
10	Spain	535	13,270	24.80	1,288

3.4 Journal analysis

A total of 13,609 articles were published in 3,973 journals, with 573 journals publishing more than 5 articles each. The top 10 journals published 1,397 articles, representing 10.3% of the total. IEEE Access led in both publication count (321 articles) and citations (14,842), significantly outperforming other journals. Future Generation Computer Systems-The International Journal of eScience, which had the highest impact factor in 2023 (6.2), ranked tenth among the top 10 journals with the most publications (Supplementary Table S1). A network visualization map, illustrating the bibliometric coupling of journals with a minimum contribution of five articles, is presented in Supplementary Figure S1. PLoS One (8,186 co-citations), IEEE Access (7,621 co-citations), and Nature (6,721 co-citations) were the three most frequently co-cited journals (Supplementary Table S2). The map illustrates the distribution of 158 highly co-cited journals, each with a minimum of 600 co-citations, serving as a hotspot map for co-cited journals (Supplementary Figure S2).

3.5 Analysis of authors

The field included 55,858 authors and 337,287 co-cited authors. Table 3 presents the top 10 most productive, highly cited, and co-cited authors. Anthony Vipin Das authored the most articles (22), followed by Wei Wang (23), and finally M. Shamin Hossain and Kyungyong Chung, each with 22. Twenty-four research clusters were identified when mapping research networks for 595 authors who had produced at least five documents (Supplementary Figure S3). Khoshgoftaar, Taghi M. received the highest number of citations (6,247), significantly more than others. The World Health Organization (1,192) was the most frequently co-cited author, followed by Chen, M. (792) and Zhang, Y. (653). Author co-citation maps were created using 1,695 authors, each cited more than 35 times (Supplementary Figure S4).

3.6 Analysis of cited references

A total of 613,873 references were found, of which 1,153 were cited more than 20 times. We displayed the top 10 articles with the

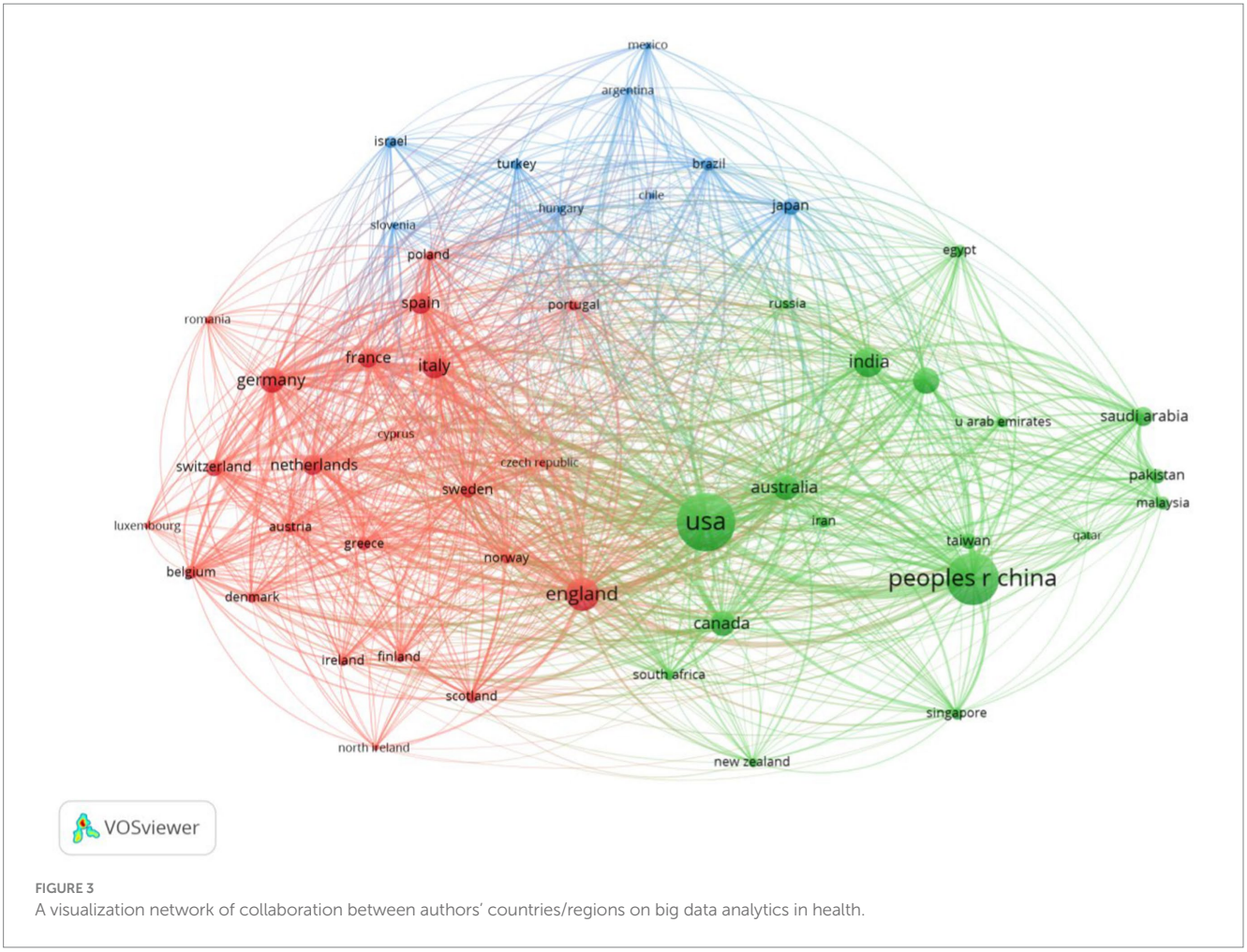

FIGURE 3
A visualization network of collaboration between authors' countries/regions on big data analytics in health.

TABLE 2 The main institutions contributing to publications on big data analytic in health.

Rank	Institutions	Country/ Region	Counts	Total citation	Average citation	Total link strength
1	Harvard Medical School	USA	211	9,011	42.71	755
2	Chinese Academy of Sciences	China	209	9,921	47.47	428
3	Stanford University	USA	169	7,636	45.18	477
4	University of Oxford	England	168	9,545	56.82	460
5	University of Toronto	Canada	167	7,572	45.34	602
6	University of Michigan	USA	160	5,629	35.18	465
7	University of California, San Francisco	USA	124	5,811	46.86	499
8	University of Pennsylvania	USA	123	5,504	44.75	487
9	University of California, Los Angeles	USA	121	3,493	28.87	354
10	University of Melbourne	Australia	119	4,456	37.45	366

highest number of citations. The top co-cited reference is by Raghupathi W. with 440 citations, followed by Murdoch T. B. with 342 citations, and Obermeyer Z. with 323 citations (Supplementary Table S3). As shown in Supplementary Figure S5, the 25 references with the greatest number of citation bursts date back to 2014. Significantly, seven citations remain in burst mode even today.

3.7 Co-occurrence analysis and clustering analysis of keywords

A total of 3,536 keywords were identified, with a cumulative frequency of 40,467 instances. Table 4 displays the 20 most frequently occurring keywords, each with an occurrence frequency of more than 450. IA keyword co-occurrence diagram was created in this domain

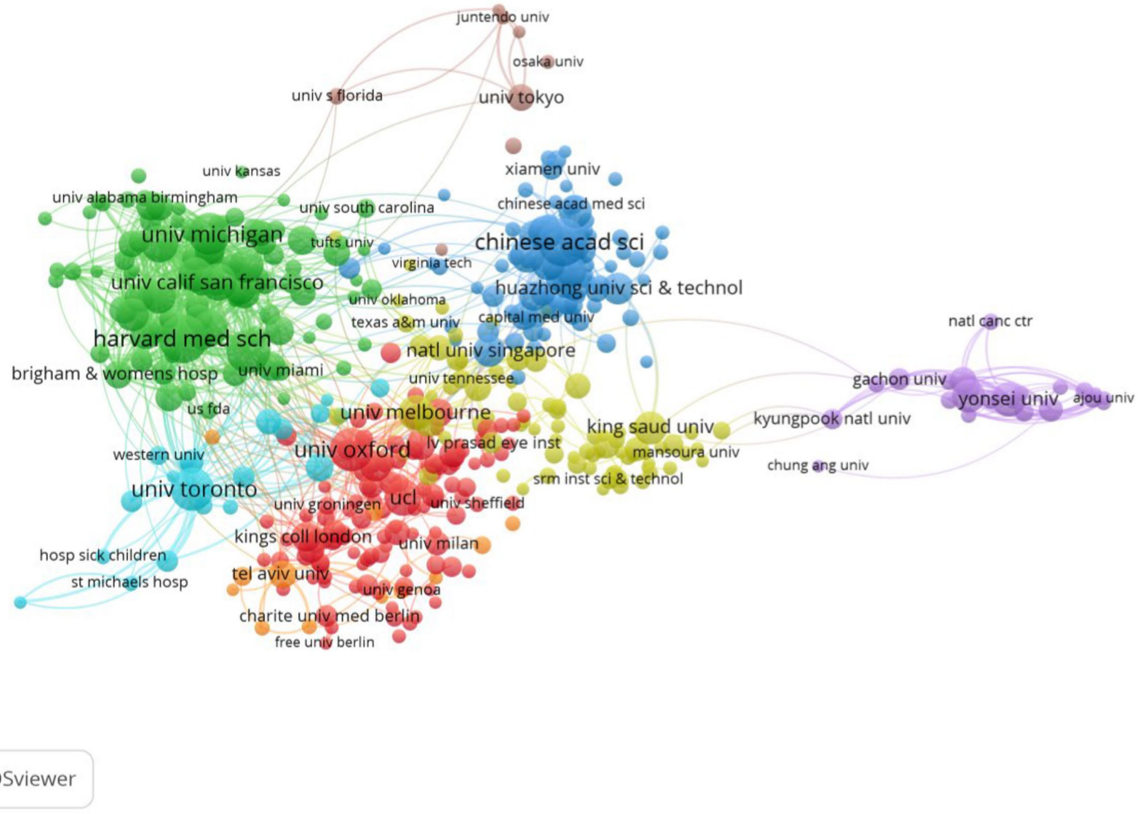

FIGURE 4
A visualization network of collaboration between authors' institutions on big data analytics in health.

TABLE 3 Top 10 productive authors and co-cited authors in the field of big data analytics in health.

Rank	Author	Counts	Cited author	Citations	Co-cited author	Citations
1	Das, Anthony Vipin	37	Khoshgoftaar, Taghi M.	6,247	World Health Organization	1,192
2	Wang, Wei	23	Chen, Min	2,018	Chen, M.	792
3	Chung, Kyungyong	22	Wang, Peng	1,679	Zhang, Y.	653
4	Hossain, M. Shamim	22	Guizani, Mohsen	1,626	Liu, Y.	623
5	Bragazzi, Nicola Luigi	21	Hossain, M. Shamin	1,433	Wang, Y	550
6	Wang, Lei	21	Zhang, Zhongheng	1,392	Obermeyer, Z.	491
7	Chen, Bin	20	Gadekallu, Thippa Reddy	1,384	Raghupathi, W.	473
8	Rodrigues, Joel J. P. C.	19	Yang, Laurence T.	1,364	Lee, J.	464
9	Wang, Hao	18	Muhammad, Ghulam	1,311	Li, Y.	441
10	Kumar, Neeraj; Li, Li; Zhang, Lei	16	Chen, Bin	1,310	Kim, J.	428

using VOSviewer, highlighting 552 keywords that appeared more than 30 times for visualization (Figure 5). Each color in the diagram represents a cluster indicating a similar subject among the publications, and each keyword is represented by a circle. The keywords were categorized into five clusters: Cluster 1 (light blue) encompassed healthcare, management, big data analytics, and performance; Cluster 2 (purple) centered on systems, the internet, cloud computing, healthcare, and challenges; Cluster 3 (yellow) highlighted machine learning, artificial intelligence, deep learning, prediction, and classification; Cluster 4 (dark blue) encompassed big data, information, privacy, and ethics; Cluster 5 (green) covered precision medicine, identification, and diagnosis; Cluster 6 (red) centered on big data analytics, risk, and mortality. The thematic terms for keyword clustering are presented in Table 5. VOSviewer color-coded the keywords on the map according to their average year of appearance (Supplementary Figure S6). The prominence of artificial intelligence and deep learning in medical big data analytics is underscored by their frequent mention in recent discussions.

TABLE 4 High-frequency keywords in the studies of big data analytics in health.

Rank	Keywords	Frequency	Total link strength
1	big data	6,170	28,869
2	machine learning	1,671	9,408
3	artificial intelligence	1,287	7,487
4	health	1,046	5,148
5	risk	700	3,702
6	deep learning	693	3,811
7	prediction	693	4,296
8	classification	678	3,910
9	COVID-19	607	2,810
10	challenges	606	4,029
11	internet	601	4,248
12	management	590	3,582
13	care	582	3,364
14	system	567	3,247
15	model	565	3,130
16	health-care	546	3,214
17	framework	493	3,331
18	big data analytics	479	2,584
19	diagnosis	458	2,635
20	precision medicine	454	2,468

3.8 Burst of keywords

Burst keywords are terms that frequently emerge over a period of time. Figure 6 highlights the top 25 keywords with the most significant citation bursts. The keyword "personalized medicine" had the highest burst value (n = 17.96), while "decision support" and "data protection" exhibited the longest burst periods. The keywords that experienced the most significant bursts from 2022 to 2024 included Industry 4.0, strategy, digital transformation, surgery, sentiment analysis, and burden.

4 Discussion

This study employed two prominent bibliometric analysis tools, VOSviewer and CiteSpace, to examine the evolution of big data research in the health sector over the past 15 years. Our analysis provided a thorough overview of the current landscape, developmental trends, and emerging research hotspots, providing valuable insights for researchers to enhance their understanding of the dynamics and future prospects of this field. The findings reveal that 71.8% of the published literature in the past 5 years underscores the rapid growth of big data applications in the medical domain. The reasons for this rapid growth may include the introduction of policies by various countries and regions to promote digital health and precision medicine, the rapid development of big data analytics technologies, and the continuous deepening of interdisciplinary collaboration. Furthermore, the outbreak of COVID-19 in 2020 prompted a substantial volume of literature focusing on how to leverage big data technologies to address pandemic challenges, such as real-time epidemic data analysis, vaccine distribution, and public health decision-making support. However, it should be noted that the expansion of the Web of Science Core Collection may have a potential impact on the growth of publication numbers (24).

The USA occupies a leading position in global health big data research, as evidenced by 4,053 published papers and a total citation count of 159,464, underscoring the country's prominent role in advancing this field. Additionally, the average citation count of USA research ranks third globally, indicating that the country excels not only in volume but also in academic impact. This trend is further supported by the USA's TLS of 3,957, highlighting its significant role in international collaborations. In comparison, China and England closely follow, with 3,184 and 1,341 publications, respectively. Notably, the average citation rate of 40.06 for England reflects the high global reputation of its research.

Through the analysis of international collaboration, we found that the USA, England, and China are core participants in the global network, a fact further supported by their strong TLS values. The collaboration network diagram generated by VOSviewer shows that these three countries occupy central positions with close cooperation ties. The increasing collaborative efforts among nations further highlight the globalization of health big data research, indicating substantial cross-border cooperation in addressing common healthcare challenges. At the institutional level, Harvard Medical School, a prestigious medical school within Harvard University, stands as a key hub in this field, having published 211 papers and leading in centrality indicators such as link strength and intermediary centrality. Six top institutions from the USA rank among the top 10, further emphasizing the country's dominant position in health big data research. The Chinese Academy of Sciences and Stanford University follow closely in terms of publication volume, showcasing the global influence of these institutions. The TLS values of leading institutions, such as Harvard Medical School (755), the University of Toronto (602), and the University of California, San Francisco (499), demonstrate their active engagement in interdisciplinary collaborations, particularly in research areas requiring large-scale data analysis. This underscores the growing significance of big data in the healthcare sector.

We observed that the top three journals have published more than 150 articles each, with impact factors exceeding 3, and are classified as Q1 or Q2 in the Journal Citation Reports (JCR) rankings. This not only indicates the high quality of research and academic influence of these journals in the field of health big data, but also reflects researchers' recognition of these journals as core platforms for knowledge dissemination. Consequently, future research should consider submitting to these high-impact journals to enhance the visibility and citation potential of their studies.

The 10 most-cited articles illuminate the primary research directions in health big data analytics, emphasizing the introduction of methodologies and risk identification. For instance, Raghupathi W. systematically discusses the structure, implementation strategies, and challenges of healthcare big data in his highly co-cited paper, "Big data analytics in healthcare: promise and potential" (23), which lays the groundwork for subsequent research in the field. Similarly, Murdoch T. B.'s review in The Journal of the American Medical Association highlights the significance and inevitability of big data

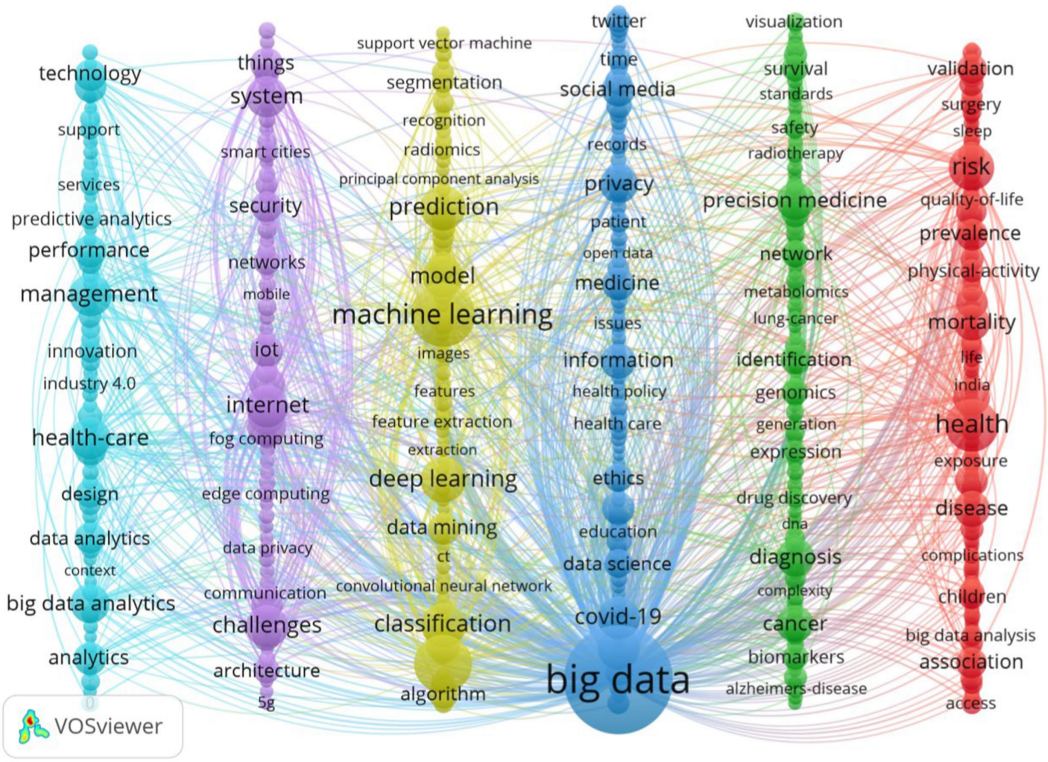

FIGURE 5
Co-occurrence mapping of keywords.

technologies in healthcare (2013) (25). These highly cited papers are predominantly review articles that examine the methodological frameworks of big data in healthcare, while high-quality original research on practical applications remains relatively scarce. Co-citation analysis reveals the dynamic evolution of research frontiers. Based on the temporal analysis of the 25 most co-cited papers, early research focused on fundamental methods for big data in medicine, including data collection, storage, and processing techniques. However, recent research trends have gradually shifted toward application areas such as disease risk identification and precision medicine (26, 27). Notably, the latest research has concentrated on the integration of big data technologies with artificial intelligence and machine learning algorithms, as well as improvements in data privacy protection measures (28, 29). This trend reflects the significant attention from both academia and industry toward the potential of big data technologies in healthcare. Furthermore, an author analysis reveals that prolific authors such as Anthony Vipin Das and Wei Wang have played crucial roles in advancing the field of health big data, while Taghi M. Khoshgoftaar has demonstrated substantial academic influence through his high citation count. Additionally, the World Health Organization, as a highly co-cited author, further highlights its central role in policy development and the application of health data.

The results from topic clustering indicate that research hotspots in healthcare big data analytics are characterized by significant diversity and interdisciplinary features. For instance, Cluster 1 emphasizes healthcare management and decision support, demonstrating the potential of big data to optimize resource allocation and improve service efficiency. In contrast, Cluster 3 centers on artificial intelligence technologies, such as machine learning and deep learning, underscoring their critical roles in health monitoring, disease prediction, and classification. With ongoing advancements in medical devices, internet connectivity, and cloud computing, vast amounts of heterogeneous big data are generated daily within the healthcare sector (30). Transforming these vast data resources into actionable knowledge bases to improve the efficiency and quality of healthcare services has become a crucial task for professionals in the field. By employing advanced data modeling techniques, such as artificial intelligence, machine learning, and deep learning, it is possible to effectively identify risk factors for patients, thereby enabling the prediction and management of high-risk populations (4). With the increasing availability of genomic data, these technologies offer a solid scientific foundation for early disease prediction, precision treatment, and personalized healthcare. For example, big data technologies have shown significant potential and value in the diagnosis and risk prediction of cardiovascular diseases (31), respiratory diseases (32), sepsis (33), and cancer (34). By enhancing the analytical capabilities of unstructured data and leveraging advanced big data technologies, clinical decision support systems can thoroughly analyze medical imaging data and extract key information from medical literature, thus constructing comprehensive medical expert knowledge bases. This technological support not only aids physicians in making more accurate diagnoses but also provides recommendations for drug dosage adjustments and personalized treatment plans.

The keywords “privacy” and “ethics” in Cluster 4 underscore the growing concerns regarding data security and privacy protection as healthcare data application expands. This suggests that the use of big data in the healthcare sector encounters numerous challenges (35). The primary topic clustering in Cluster 2, which pertains to the

TABLE 5 Detailed list of clustered keywords and thematic terms.

Cluster	Theme	Representative keywords
#1	Application of big data analytics in health decision-making	Health care; management; decision making; technology; support; supply chain; services; innovation; predictive analytics; performance
#2	Challenges in the technological management of health and medical big data	Internet; IoT; fog computing; edge computing; 5G; mobile; systems; challenges; data privacy; big data applications; data management; medical big data
#3	Integration of machine learning with health monitoring	Machine learning; deep learning; data mining; algorithm; feature extraction; model; prediction; classification; prognostics; health monitoring
#4	Privacy and ethical issues in health and medical big data	Big data; data science; social media; health information; electronic health record; digital health; patient; medicine; ethics; health policy; data protection
#5	Data integration and application in precision medicine	Data integration; precision medicine; treatment; diagnosis; prognosis; survival; cancer; Alzheimer's disease; genomics; DNA; biomarkers
#6	Application of big data in disease management and risk assessment	Big data analysis; health; disease; administrative data; prevention; risk; prevalence; exposure; complications

challenges of implementing big data technologies such as the Internet of Things (IoT) in healthcare, further reinforces this assertion. A major challenge lies in ensuring the security of healthcare information, particularly when it involves sensitive patient data and personal privacy (36). Moreover, despite the abundance of available healthcare data, effectively integrating and improving the overall quality of this data remains a considerable challenge (37). Currently, there is a shortage of professionals who possess both medical expertise and advanced big data analysis skills, including machine learning, artificial intelligence, and deep learning. This deficiency is insufficient to meet the health needs of the broader population. Consequently, this relative shortage of talent restricts the efficiency of data management and somewhat hinders the effectiveness of data sharing (22). Therefore, it is imperative to continue accelerating the training of interdisciplinary experts who possess both medical backgrounds and advanced big data analysis skills (38).

The keyword color distribution over time indicates a gradual shift in recent research trends towards areas like artificial intelligence and precision medicine. For instance, the higher average values of keywords like "artificial intelligence" and "deep learning" suggest that these technologies have emerged as key research hotspots in healthcare big data in recent years. Concurrently, the keywords "precision medicine" and "diagnosis" in Cluster 5 reflect the healthcare sector's accelerating transition from traditional broad-spectrum treatments to personalized and precision-based approaches, supported by big data-related technologies. Notably, during the global COVID-19 pandemic, big data management and analysis technologies played a critical role, demonstrating their vast potential in infectious disease monitoring and control (39). These technologies are essential for the rapid detection and prevention of infectious disease spread, further underscoring the importance of big data in public health emergency management. Additionally, the persistent emergence of keywords like "decision support" and "data protection" highlights the ongoing importance of these themes in healthcare big data analytics, aligning with the increasing trend of utilizing healthcare big data in real-time decision support systems. From 2022 to 2024, the notable rise of keywords like "Industry 4.0," "digital transformation," and "sentiment analysis" suggests that the healthcare sector is accelerating its integration with other cutting-edge technologies, including IoT and natural language processing, which are crucial for advancing the development of the broader health industry. Furthermore, the development of "smart healthcare" or "digital health" focuses on leveraging technology to enhance medical services, improve patient care, and enable real-time health monitoring. The theoretical frameworks of "data integration," "data collaboration," and others have become key components in the application of big data in health management, ensuring that data from various health domains can seamlessly work together (40). Overall, the rise of hot topics like artificial intelligence, precision medicine, digital transformation and privacy ethics reflects both the technology-driven changes at the academic frontier and the challenges and needs faced by the healthcare sector in practical applications. However, there are certain limitations in the current research, including the insufficient exploration of regional differences in research topics. Future studies could further investigate the diversification and localization characteristics of global healthcare big data applications. Moreover, how to balance the ethical risks of big data technologies with their societal benefits is an important direction that requires in-depth consideration.

This study updates the analysis of the most recently published articles to identify the current research trends in the application of big data in the healthcare sector. Before conducting the literature search, we reviewed several high-quality research papers, carefully extracted relevant keywords, and extensively consulted with domain experts to develop a comprehensive and scientifically sound search strategy. As a result, our search method is both rigorous and systematic, ensuring the accuracy and reliability of the findings. However, there are some inherent limitations in this bibliometric study. First, our analysis was limited to English-language articles from the WoS Core Collection database, which may have resulted in the exclusion of significant literature published in other languages or in journals not indexed by WoS Core Collection, potentially affecting the comprehensiveness of the study. This limitation is common in bibliometric research, primarily due to the incompatibility between databases when conducting comparative analyses, particularly regarding bibliometric impact metrics. These differences are largely attributable to the varying coverage of journals, conference proceedings, and books across different databases. Therefore, to ensure analytical consistency and facilitate a broader exploration of data fields and metadata, we confined our analysis to a specific database, the WoS Core Collection. Second, recent publications, particularly those from 2024, typically have fewer citations, as they may not yet have gained sufficient recognition yet. This could delay the identification of the latest research advancements. Moreover, citation counts can be influenced by various

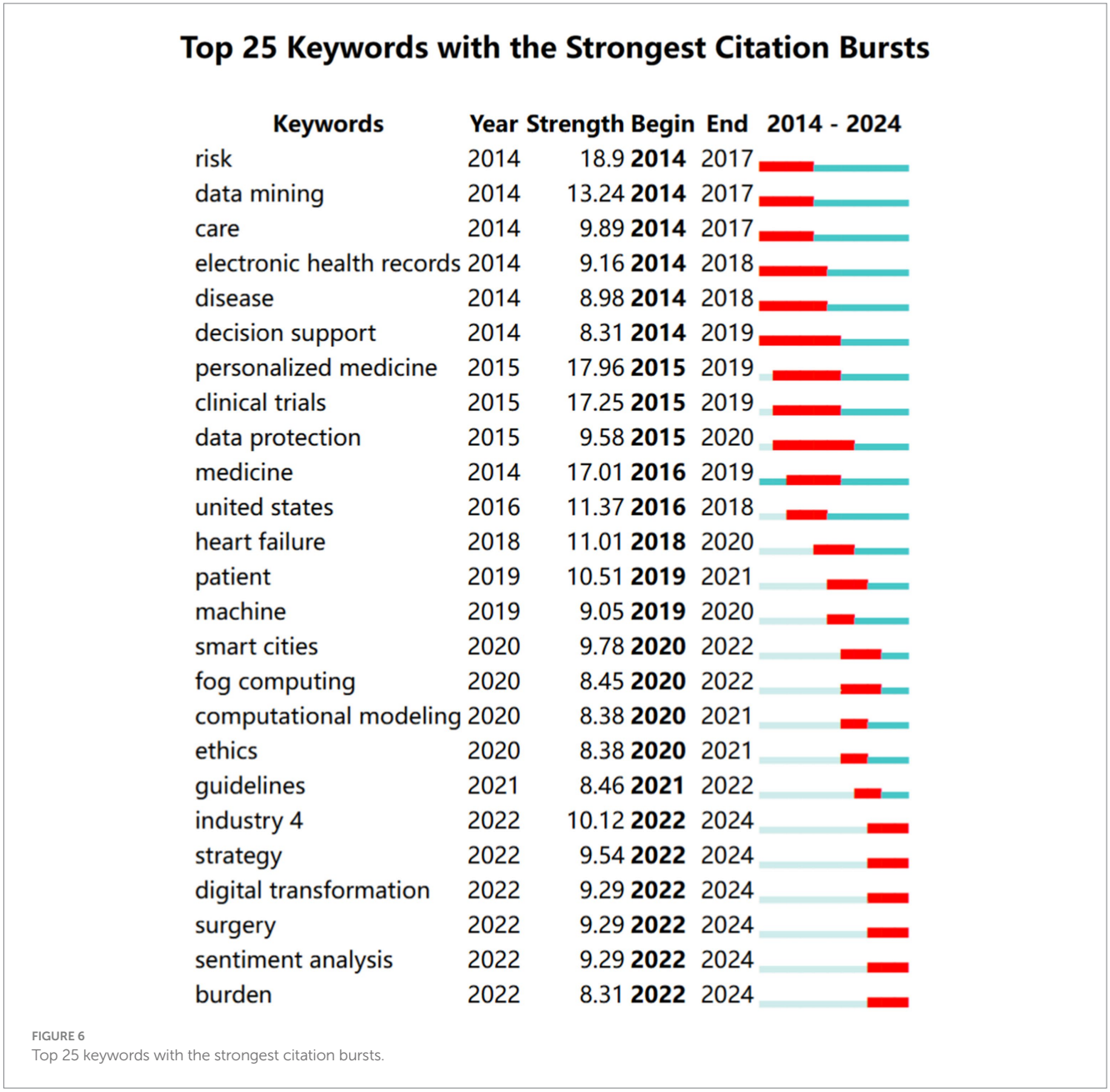

FIGURE 6
Top 25 keywords with the strongest citation bursts.

factors, including journal impact factors, author self-citations, incomplete references, and citation biases. As a result, some groundbreaking studies may initially receive limited citations until their significance is more broadly acknowledged. This phenomenon, known in bibliometrics as "forgotten through absorption," has been widely discussed and recognized in previous research. Meanwhile, it is important to note that the WoS Core Collection has certain limitations when conducting historical literature retrieval. According to Liu (41), the WoS Core Collection faces challenges in retrieving older literature, including issues such as limitations in topic searches. As a result, the relatively low number of early publications in this study may be closely associated with these technical constraints. Finally, the inability to conduct a comprehensive funding analysis constitutes another limitation of this study. Future research could investigate this aspect further, thereby offering additional insights into the influence of financial support on the direction of research. Moreover, although bibliometric analysis is highly useful for identifying academic trends, it may not fully capture the long-term impact of big data in the healthcare sector. To address this issue, supplementary methods such as case studies, policy evaluations, and technology assessments could be considered to gain a deeper understanding of the enduring effects of big data in this field. Future research could further assess the sustained impact of big data technologies in healthcare through longitudinal studies.

5 Conclusion

We analyzed the application of big data in the health sector over the past 15 years using bibliometric tools including VOSviewer and CiteSpace. Our research indicates that big data technologies have substantial potential to enhance the accuracy of disease diagnostics and treatment outcomes, particularly in the realms of cancer treatment and disease risk assessment. However, we also identified several limitations within big data research, including constraints related to

data sources and challenges in data processing. With advancements in data analysis technologies and the refinement of healthcare policies, big data is expected to play a more significant role in public health management, disease prevention, and health promotion. We recommend prioritizing interdisciplinary collaboration to integrate and apply big data technologies in healthcare, optimizing medical services, and improving public health outcomes.

Author contributions

LY: Writing – original draft, Writing – review & editing, Data curation, Methodology, Formal analysis, Visualization, Software. YL: Writing – original draft, Writing – review & editing, Data curation, Methodology, Formal analysis, Visualization, Software. TW: Writing – review & editing, Methodology, Formal analysis, Software. CH: Writing – review & editing, Methodology, Formal analysis, Funding acquisition. QxL: Writing – review & editing, Conceptualization, Funding acquisition. QL: Writing – review & editing. XY: Writing – review & editing. TR: Writing – review & editing. YW: Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This research was supported by the Affiliated Hospital of Guizhou Medical University National Natural Science Foundation Cultivation Project (grant no. GYFYNSFC[2023-35]).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmed.2025.1456286/full#supplementary-material

References

1. Rodríguez-Mazahua L, Rodríguez-Enríquez C-A, Sánchez-Cervantes JL, Cervantes J, García-Alcaraz JL, Alor-Hernández G. A general perspective of big data: applications, tools, challenges and trends. *J Supercomput*. (2016) 72:3073–113. doi: 10.1007/s11227-015-1501-1

2. Xia S, Song J, Ameen N, Vrontis D, Yan J, Chen F. What changes and opportunities does big data analytics capability bring to strategic alliance research? A systematic literature review. *Int J Manag Rev*. (2024) 26:34–53. doi: 10.1111/ijmr.12350

3. Han X, Gstrein OJ, Andrikopoulos V. When we talk about big data, what do we really mean? Toward a more precise definition of big data. *Front Big Data*. (2024) 7:1441869. doi: 10.3389/fdata.2024.1441869

4. Schulte T, Bohnet-Joschko S. How can big data analytics support people-centred and integrated health services: a scoping review. *Int J Integr Care*. (2022) 22:23. doi: 10.5334/ijic.5543

5. Obermeyer Z, Emanuel EJ. Predicting the future—big data, machine learning, and clinical medicine. *N Engl J Med*. (2016) 375:1216–9. doi: 10.1056/NEJMp1606181

6. Galetsi P, Katsaliaki K. Big data analytics in health: an overview and bibliometric study of research activity. *Health Inf Libr J*. (2020) 37:5–25. doi: 10.1111/hir.12286

7. Lorenzo G, Ahmed SR, Hormuth DA, Vaughn B, Kalpathy-Cramer J, Solorio L, et al. Patient-specific, mechanistic models of tumor growth incorporating artificial intelligence and big data. *Annu Rev Biomed Eng*. (2024) 26:529–60. doi: 10.1146/annurev-bioeng-081623-025834

8. Dong J, Wu H, Zhou D, Li K, Zhang Y, Ji H, et al. Application of big data and artificial intelligence in COVID-19 prevention, diagnosis, treatment and management decisions in China. *J Med Syst*. (2021) 45:84. doi: 10.1007/s10916-021-01757-0

9. Kindle RD, Badawi O, Celi LA, Sturland S. Intensive care unit telemedicine in the era of big data, artificial intelligence, and computer clinical decision support systems. *Crit Care Clin*. (2019) 35:483–95. doi: 10.1016/j.ccc.2019.02.005

10. Ellegaard O, Wallin JA. The bibliometric analysis of scholarly production: how great is the impact? *Scientometrics*. (2015) 105:1809–31. doi: 10.1007/s11192-015-1645-z

11. Van Eck NJ, Waltman L. Citation-based clustering of publications using CitNetExplorer and VOSviewer. *Scientometrics*. (2017) 111:1053–70. doi: 10.1007/s11192-017-2300-7

12. Sgrò A, Al-Busaidi IS, Wells CI, Vervoort D, Venturini S, Farina V, et al. Global surgery: a 30-year bibliometric analysis (1987–2017). *World J Surg*. (2019) 43:2689–98. doi: 10.1007/s00268-019-05112-w

13. Shen Z, Hu J, Wu H, Chen Z, Wu W, Lin J, et al. Global research trends and foci of artificial intelligence-based tumor pathology: a scientometric study. *J Transl Med*. (2022) 20:409. doi: 10.1186/s12967-022-03615-0

14. Wang Y, Liu Q, Chen Y, Qian Y, Pan B, Ge L, et al. Global trends and future prospects of child nutrition: a bibliometric analysis of highly cited papers. *Front Pediatr*. (2021) 9:633525. doi: 10.3389/fped.2021.633525

15. Amusa LB, Twinomurinzi H, Phalane E, Phaswana-Mafuya RN. Big data and infectious disease epidemiology: bibliometric analysis and research agenda. *Interact J Med Res*. (2023) 12:e42292. doi: 10.2196/42292

16. Liang C, Qiao S, Olatosi B, Lyu T, Li X. Emergence and evolution of big data science in HIV research: bibliometric analysis of federally sponsored studies 2000–2019. *Int J Med Inform*. (2021) 154:104558. doi: 10.1016/j.ijmedinf.2021.104558

17. Cui X, Chang Y, Yang C, Cong Z, Wang B, Leng Y. Development and trends in artificial intelligence in critical care medicine: a bibliometric analysis of related research over the period of 2010–2021. *J Pers Med*. (2022) 13:50. doi: 10.3390/jpm13010050

18. Gu D, Li J, Li X, Liang C. Visualizing the knowledge structure and evolution of big data research in healthcare informatics. *Int J Med Inform*. (2017) 98:22–32. doi: 10.1016/j.ijmedinf.2016.11.006

19. Subrahmanya SVG, Shetty DK, Patil V, Hameed BMZ, Paul R, Smriti K, et al. The role of data science in healthcare advancements: applications, benefits, and future prospects. *Ir J Med Sci*. (2022) 191:1473–83. doi: 10.1007/s11845-021-02730-z

20. Aria M, Cuccurullo C. *bibliometrix*: An R-tool for comprehensive science mapping analysis. *J Informetr*. (2017) 11:959–75. doi: 10.1016/j.joi.2017.08.007

21. Li K, Rollins J, Yan E. Web of science use in published research and review papers 1997–2017: a selective, dynamic, cross-domain, content-based analysis. *Scientometrics*. (2018) 115:1–20. doi: 10.1007/s11192-017-2622-5

22. Saenz A, Chen E, Marklund H, Rajpurkar P. The MAIDA initiative: establishing a framework for global medical-imaging data sharing. *Lancet Digit Health*. (2024) 6:e6–8. doi: 10.1016/S2589-7500(23)00222-4

23. Raghupathi W, Raghupathi V. Big data analytics in healthcare: promise and potential. *Health Inf Sci Syst*. (2014) 2:3. doi: 10.1186/2047-2501-2-3

24. Liu W, Ni R, Hu G. Web of Science Core Collection's coverage expansion: the forgotten Arts & Humanities Citation Index. *Scientometrics*. (2024) 129:933–55. doi: 10.1007/s11192-023-04917-w

25. Murdoch TB, Detsky AS. The inevitable application of big data to health care. *JAMA*. (2013) 309:1351–2. doi: 10.1001/jama.2013.393

26. Bates DW, Saria S, Ohno-Machado L, Shah A, Escobar G. Big data in health care: using analytics to identify and manage high-risk and high-cost patients. *Health Aff*. (2014) 33:1123–31. doi: 10.1377/hlthaff.2014.0041

27. Collins FS, Varmus H. A new initiative on precision medicine. *N Engl J Med*. (2015) 372:793–5. doi: 10.1056/NEJMp1500523

28. Ngiam KY, Khor W. Big data and machine learning algorithms for health-care delivery. *Lancet Oncol*. (2019) 20:e262–73. doi: 10.1016/S1470-2045(19)30149-4

29. Price WN, Cohen IG. Privacy in the age of medical big data. *Nat Med*. (2019) 25:37–43. doi: 10.1038/s41591-018-0272-7

30. Ismail L, Materwala H, Karduck AP, Adem A. Requirements of health data management systems for biomedical care and research: scoping review. *J Med Internet Res*. (2020) 22:e17508. doi: 10.2196/17508

31. Xie P, Wang H, Xiao J, Xu F, Liu J, Chen Z, et al. Development and validation of an explainable deep learning model to predict in-hospital mortality for patients with acute myocardial infarction: algorithm development and validation study. *J Med Internet Res*. (2024) 26:e49848. doi: 10.2196/49848

32. Pereira AM, Jácome C, Jacinto T, Amaral R, Pereira M, Sá-Sousa A, et al. Multidisciplinary development and initial validation of a clinical knowledge base on chronic respiratory diseases for mHealth decision support systems. *J Med Internet Res*. (2023) 25:e45364. doi: 10.2196/45364

33. Su L, Xu Z, Chang F, Ma Y, Liu S, Jiang H, et al. Early prediction of mortality, severity, and length of stay in the intensive care unit of sepsis patients based on sepsis 3.0 by machine learning models. *Front Med*. (2021) 8:664966. doi: 10.3389/fmed.2021.664966

34. Xiang H, Xiao Y, Li F, Li C, Liu L, Deng T, et al. Development and validation of an interpretable model integrating multimodal information for improving ovarian cancer diagnosis. *Nat Commun*. (2024) 15:2681. doi: 10.1038/s41467-024-46700-2

35. Yang X, Huang K, Yang D, Zhao W, Zhou X. Biomedical big data technologies, applications, and challenges for precision medicine: a review. *Global Chall*. (2024) 8:2300163. doi: 10.1002/gch2.202300163

36. Pool J, Akhlaghpour S, Fatehi F, Burton-Jones A. A systematic analysis of failures in protecting personal health data: a scoping review. *Int J Inf Manag*. (2024) 74:102719. doi: 10.1016/j.ijinfomgt.2023.102719

37. Alashlam L, Alzubi A. Taxonomic exploration of healthcare IoT: challenges, solutions, and future frontiers. *Appl Sci*. (2023) 13:12135. doi: 10.3390/app132212135

38. Seth P, Hueppchen N, Miller SD, Rudzicz F, Ding J, Parakh K, et al. Data science as a core competency in undergraduate medical education in the age of artificial intelligence in health care. *JMIR Med Educ*. (2023) 9:e46344. doi: 10.2196/46344

39. Mahboub B, Bataineh MTA, Alshraideh H, Hamoudi R, Salameh L, Shamayleh A. Prediction of COVID-19 hospital length of stay and risk of death using artificial intelligence-based modeling. *Front Med*. (2021) 8:592336. doi: 10.3389/fmed.2021.592336

40. Sarkar M, Lee T-H, Sahoo PK. Smart healthcare: exploring the internet of medical things with ambient intelligence. *Electronics*. (2024) 13:2309. doi: 10.3390/electronics13122309

41. Liu W. Caveats for the use of Web of Science Core collection in old literature retrieval and historical bibliometric analysis. *Technol Forecast Soc Chang*. (2021) 172:121023. doi: 10.1016/j.techfore.2021.121023