Review Article

DIGITAL HEALTH

DIGITAL HEALTH
Volume 11: 1–17
© The Author(s) 2025
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/20552076251320302
journals.sagepub.com/home/dhj
Sage

Data governance and open sharing in the fields of life sciences and medicine: A bibliometric analysis

Yanrui Qiu and Zhimin Hu

Abstract

Objective: This study aims to conduct a bibliometric analysis of literature related to data governance and open sharing in the fields of life sciences and medicine, so as to clarify the characteristics of publications and explore research hotspots and trends.

Methods: A total of 2529 valid documents published in the Web of Science Core Collection database from 2000 to 2024 were included in this study. VOSviewer was used for co-occurrence analysis, while CiteSpace was employed for clustering, burst detection, and thematic evolution analysis.

Results: Between 2000 and 2024, the number of studies on data governance and open sharing in the fields of life sciences and medicine has increased annually, indicating the growing importance of research in this area. The USA led as the country with the most research output in this field. The University of Oxford was the institution with the highest publication volume, Amy L. McGuire was the most active author, and the *Journal of Medical Internet Research* and the *Journal of the American Medical Informatics Association* were the most frequent publication outlets. The most cited reference was 'Comment: The FAIR Guiding Principles for Scientific Data Management and Stewardship'.

Conclusions: Topics such as the FAIR principles, ethical issues, public attitudes toward data sharing, data quality, databases, and big data analysis techniques are hotspots in this field. Potential research frontiers include the FAIR principles, data quality, public trust and attitudes toward data sharing, the application of artificial intelligence technology in data governance and sharing, governance and sharing of epidemiological and public health data, governance and sharing of data on chronic diseases such as diabetes, and the construction of data governance models.

Keywords

Data governance, data open sharing, life sciences and medicine, bibliometric analysis, CiteSpace, VOSviewer

Submission date: 22 October 2024; Acceptance date: 28 January 2025

Synonyms

Headword	Synonyms
Data governance and open sharing	Data governance and sharing; data sharing and governance; governance and sharing of data; sharing and governance of data
Data sharing	Data open sharing; data openness; data opening and sharing

Introduction

In November 2021, the 41st session of the General Conference of the United Nations Educational, Scientific

School of Health Policy and Management, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing, China

Corresponding author:
Zhimin Hu, School of Health Policy and Management, Chinese Academy of Medical Sciences & Peking Union Medical College, No. 9 Dongdan Santiao, Dongcheng District, Beijing 100730, China.
Email: huzhimin@pumc.edu.cn

and Cultural Organization adopted the ‘Recommendation on Open Science’, marking a new phase in the global consensus on open science.[1] As a key element of open science, the open sharing of scientific data has become a focal point of attention for countries around the world. Current scientific research is moving toward a data-intensive, data-driven, and data-sharing direction, which poses higher demands on the volume and quality of open scientific data sharing. It also brings a series of issues and challenges, such as the ownership of data rights and the risks associated with data sharing. The FAIR principles, which stand for Findable, Accessible, Interoperable, and Reusable, have established the basic guidelines for data governance. However, how to more scientifically and normatively promote data sharing and governance still requires in-depth research and discussion.

Since the initiation of the Human Genome Project, omics technologies represented by next-generation sequencing and mass spectrometry have advanced rapidly. This has led to an exponential increase in vast amounts of life science omics data, including genomics, transcriptomics, epigenomics, proteomics, and metabolomics.[2] The fields of life sciences and medicine are experiencing a profound transformation toward a data-intensive fourth paradigm of science. The data in the field of life science and medicine are characterized by enormous scale, wide variety, complex structure, and uneven quality, which makes it difficult to achieve high-dimensional and multi-level integration and sharing, thus obscuring the potential high value of scientific data. Furthermore, there is ambiguity in the ownership of individual-level health data,[3] and there are risks associated with human genetic resource-related data sharing.[4] Both nations and citizens remain skeptical about whether to share such data, which also calls for a more compatible approach to data governance. Additionally, there is a significant amount of data concealment behavior in the field of life sciences and medicine. Although data sharing among researchers can create significant public value, the potential loss of scientific leadership and economic benefits largely hinders the sharing of valuable data.[5]

Despite the various challenges facing scientific data governance and open sharing, the scientific field is transitioning from a traditional closed research paradigm to a comprehensive open science paradigm, and the sharing of scientific data is an irreversible trend. In the 1980s and 1990s, the USA, Europe, and Japan each established one of the world’s three major biological data centers: the National Center for Biotechnology Information (NCBI), the European Bioinformatics Institute (EBI), and the DNA Data Bank of Japan (DDBJ). These three data centers have greatly facilitated life science and biomedical research by sharing data resources submitted by third parties. In addition, the Cancer Genome Atlas (TCGA) database established by the National Cancer Institute in the USA and the national cohort UK Biobank (UKB) in the UK, have implemented tiered sharing of data produced by large-scale research projects. Moreover, numerous databases and knowledge bases that provide services for data query, browsing, download, and analysis have been established by small and medium-sized research team.[2]

In an effort to encourage more extensive data sharing, numerous data sharing practices have been implemented by countries worldwide. The National Institutes of Health (NIH) began requiring data-sharing plans in 2003 for grant applications with an estimated annual cost of over $500,000. Other funding agencies and organizations, including the National Science Foundation, the Howard Hughes Medical Institute, and the Wellcome Trust, have followed suit.[5] The European Union’s Open Data and the Reuse of Public-sector Information obliges member states to implement open access policies for research data generated by public funding.[1] In China, the Scientific Data Management Policy stipulates that scientific data generated from government budgetary funds should be made available for sharing with society and relevant departments as a norm, with non-sharing being the exception.[1] On the legal front, the General Data Protection Regulation (GDPR) of the European Union, which came into effect on 25 May 2018, has established rules for the protection of individuals with regard to the processing and the free movement of personal data, significantly impacting the fields of data governance and open sharing. Subsequently, the Data Governance Act, effective as of 23 June 2022, with its scope covering both personal and non-personal data, aims to promote data sharing across sectors and EU countries to harness the potential of data for the benefit of European citizens and businesses.[6]

Scientific data are gradually evolving into a factor of production, research on data governance and open sharing is increasing day by day. As a significant domain for data generation, the field of life sciences and medicine has a research landscape in data governance and open sharing. However, the current state of data governance and open sharing in the life sciences and medical fields, the contributions of various entities, the evolution of research themes, the hot issues that scholars are most concerned about, and the future development direction of this research field are still unknown and await organization. To date, no comprehensive analysis of the published status, thematic hotspots, and evolutionary trends within this field has been identified, making it necessary to discuss and summarize the current state of affairs.

Bibliometric research provides an evidence-based quantitative analysis model to understand the knowledge structure, collaboration, and frontiers of research areas. It can reveal the collaboration between countries, institutions, and authors, perform citation counts and co-citation analysis, and analyze research hotspots and trends through keyword co-occurrence and burst detection.[7–9] Bibliometric analysis can assist scholars in understanding the current state and trends of research in a particular field, providing references and insights for future research directions, and has been widely applied across various research areas. For instance, it has been used to study the global trends related to acute kidney injury in COVID-19,[10] to explore the current state, research progress, and prospects of artificial intelligence applications in wastewater treatment,[11] and to investigate the relationship between financial development and economic growth.[12] In the realm of data-related fields, Lee and Syn conducted a bibliometric analysis of the global research trends in research data management,[13] and Pradhan and Zala performed a comparative bibliometric analysis of global research literature on research data management in the Scopus and Web of Science databases from 2000 to 2019.[14] This study applies bibliometric methods to the analysis of literature related to data governance and open sharing in the life sciences and medical fields, with the aim of clarifying the collaboration, research status, and thematic evolution, as well as providing references and insights for future research in this area.

CiteSpace and VOSviewer are two Java-based information visualization software tools. This study utilizes these two tools in combination to conduct a comprehensive bibliometric analysis of publications related to data governance and open sharing in the field of life sciences and medicine, in order to explore the characteristics of publications, thematic hotspots, evolutionary trends, and future research directions in this domain.

This study analyzes the publication quantity, co-occurrence, clustering, and burst detection from the perspectives of entities such as countries, institutions, authors, journals, references, and keywords. The research objectives of this study primarily encompass the following three aspects:

1. Identifying the leading countries, institutions, and authors in the field of data governance and open sharing in life sciences and medicine, as well as their collaboration situation.
2. Exploring the evolution and development of research themes within this field.
3. Investigating the hotspots and forecasting the frontiers of research in the field.

Methods

Data source and literature search strategy

The Web of Science is the most comprehensive academic information resource globally, covering the largest number of disciplines, with over 12,000 core academic journals included. It is frequently used by researchers and widely recognized as a reliable and comprehensive source of academic information, making it the preferred choice for conducting bibliometric analysis.[8,9] We found that using 'Topic' for the search yielded a large number of irrelevant and redundant documents, so this study employs 'Title' for literature retrieval, which not only ensures a sufficient sample size of documents but also guarantees the precision of the search. Additionally, we utilized the MeSH thesaurus of the PubMed database to search for synonyms of 'data sharing' and 'data governance', and ultimately determined the search strategy through screening. Moreover, 'data governance' is a term that has emerged in recent years, with 'data management' being more commonly used in the past. To fully present the evolution of the subject, this study included both 'data governance' and 'data management' in the search strategy. To ensure the timeliness of the research, this study retrieved literature from the Web of Science Core Collection database from 1 January 2000 to 24 March 2024. The search term was: TI = ('Open Data Sharing' OR 'Data Openness' OR 'Data Sharing' OR 'Data Governance' OR 'Data Management'). To facilitate further analysis of the literature content, the search was limited to documents in 'English' and the document type was specified as 'article'. The research areas were filtered to encompass disciplines within the life sciences and medical fields. The retrieved records were then imported into NoteExpress for deduplication, resulting in a final set of 2529 valid documents. The obtained documents were exported in 'plain text file' format, including full records and cited references.

Software for bibliometric analysis

This study primarily utilized CiteSpace 5.5.R2 and VOSviewer 1.6.18 as the tools for bibliometric analysis, with Excel and Scimago Graphica software used for visualization. Figure 1 presents the literature retrieval strategy, inclusion and exclusion criteria, and the analytical approach. VOSviewer was predominantly used for co-occurrence analysis of countries, institutions, authors, and keywords, while CiteSpace was mainly applied for clustering and burst detection analysis. Excel was employed for frequency statistics, and Scimago Graphica was used for generating geographical visualization maps.

Results

Analysis of annual publication volume

Figure 2 illustrates the distribution of annual publication volumes related to data governance and open sharing in the fields of life sciences and medicine from the WOSCC database between 2000 and 2024, along with its correlation with an exponentially growing predictive model. Since the year 2000, the annual number of publications in this field has shown a fluctuating upward trend, peaking in the year 2021, after which there was a slight decline in the annual volume of publications. An exponential growth function was used to evaluate the relationship between the annual

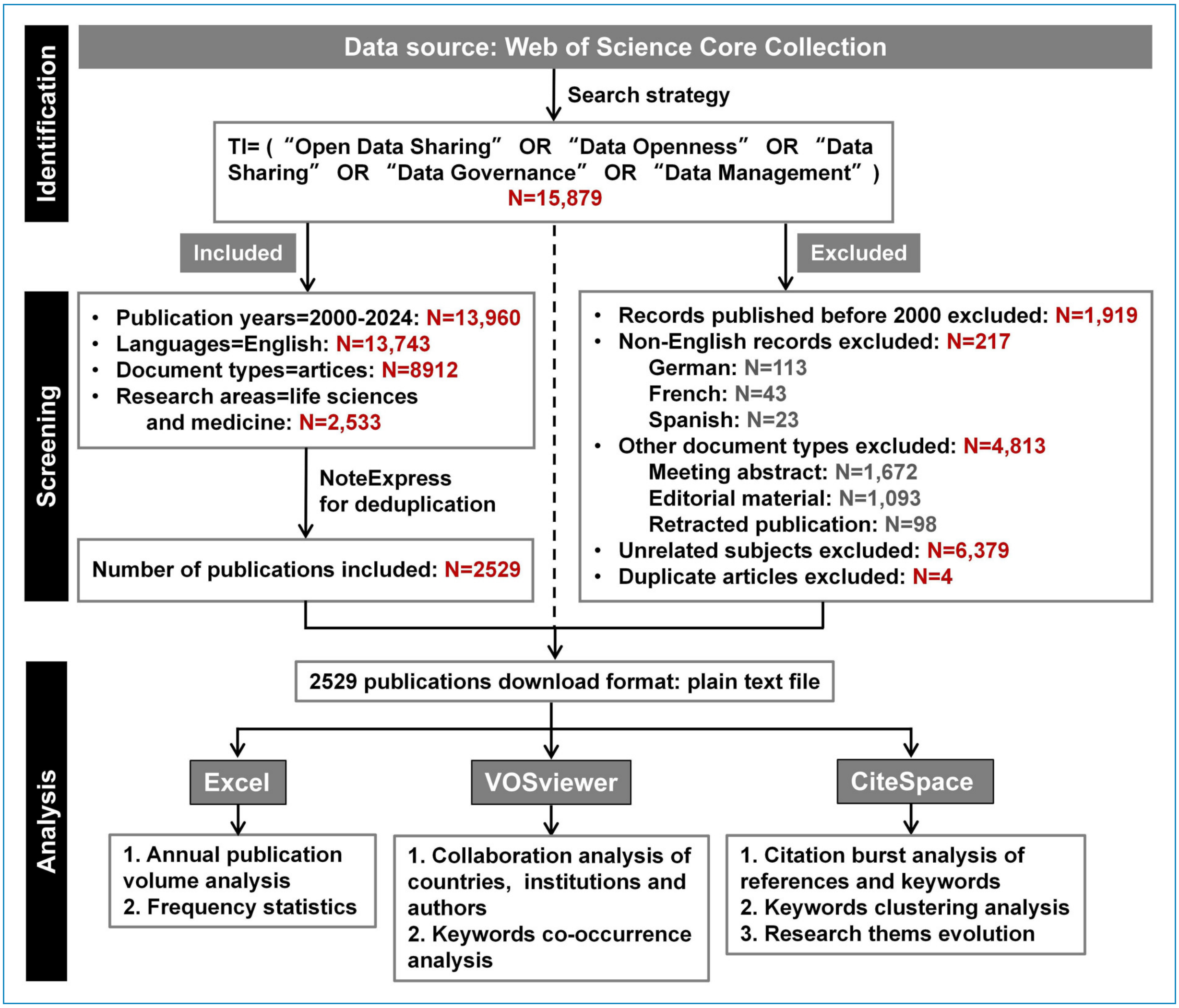

Identification
Data source: Web of Science Core Collection
Search strategy
TI= ("Open Data Sharing" OR "Data Openness" OR "Data Sharing" OR "Data Governance" OR "Data Management")
N=15,879
Included
Excluded
Screening
• Publication years=2000-2024: N=13,960
• Languages=English: N=13,743
• Document types=artices: N=8912
• Research areas=life sciences and medicine: N=2,533
NoteExpress for deduplication
Number of publications included: N=2529
• Records published before 2000 excluded: N=1,919
• Non-English records excluded: N=217
German: N=113
French: N=43
Spanish: N=23
• Other document types excluded: N=4,813
Meeting abstract: N=1,672
Editorial material: N=1,093
Retracted publication: N=98
• Unrelated subjects excluded: N=6,379
• Duplicate articles excluded: N=4
Analysis
2529 publications download format: plain text file
Excel
VOSviewer
CiteSpace
1. Annual publication volume analysis
2. Frequency statistics
1. Collaboration analysis of countries, institutions and authors
2. Keywords co-occurrence analysis
1. Citation burst analysis of references and keywords
2. Keywords clustering analysis
3. Research thems evolution

Figure 1. Flowchart of the literature retrieval strategy, inclusion and exclusion criteria, and the analytical approach.

publication volume and the year of publication. The results showed that the model was in good agreement with the observed trend in publication volumes ($R^2 = .7221$). This strong correlation indicates that research on data governance and open sharing in the life sciences and medicine has undergone significant growth and development, presenting a thriving landscape in the new century.

Analysis of national publications

This study revealed the contributions and collaboration among countries/regions by counting the number of publications and conducting geographical visualization analysis. The top 10 countries/regions by publication volume and the collaboration map are depicted in Figure 3. According to the analysis of the included literature, between 2000 and 2024, a total of 133 countries/regions participated in the publication of documents in this field, forming 15 clustered networks. The USA ranked first with 1041 publications, significantly higher than the UK in second place (435) and Germany in third (250). The publication volumes of the other countries in the top 10 were relatively close, and the countries/regions not included in Figure 3a had publication volumes that did not exceed 100 documents. Due to the phenomenon of multinational collaboration in scientific research activities, collaborative papers are counted more than once, hence the sum of the publication volumes of the countries is greater than the total number of articles included. This study will count all the countries involved in the publication of a particular literature, with the final primary statistical indicator being the absolute number of publications each country has contributed to. The statistics for institutions and authors will be similar to this approach.

The study combined VOSviewer and Scimago Graphica software to select the top 30 countries/regions by publication volume and mapped out an international collaboration network, which was then displayed using geographical visualization charts to depict the state of international collaboration

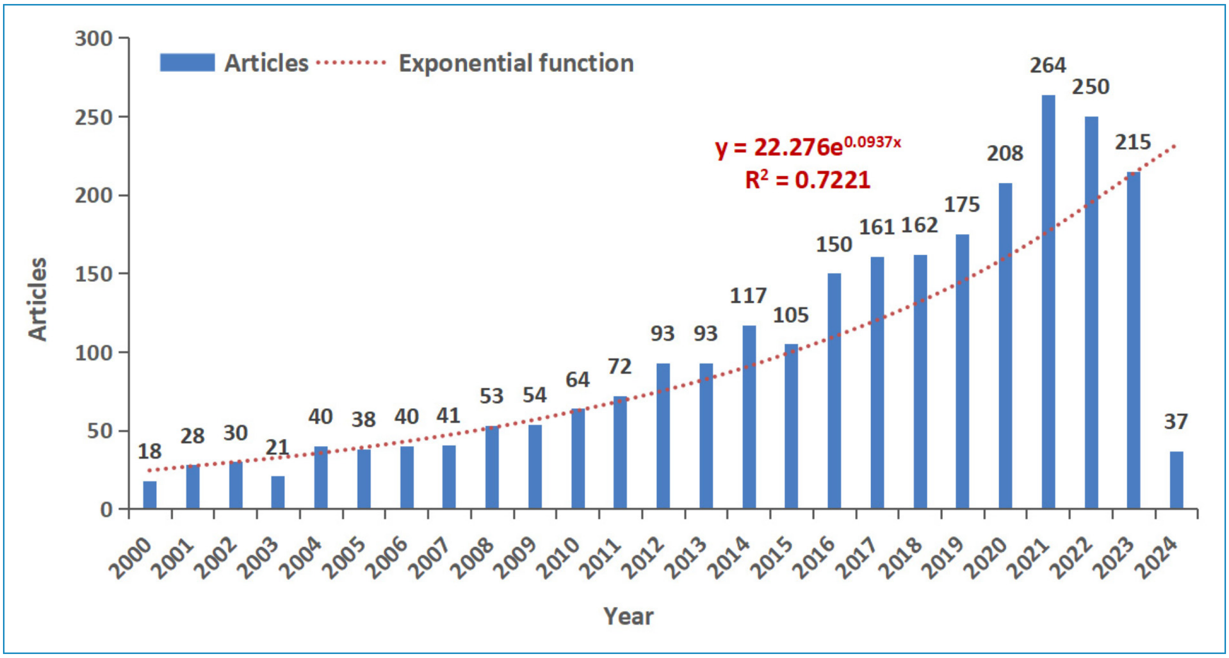


Figure 2. Annual publication volume and the exponential function predictive model.

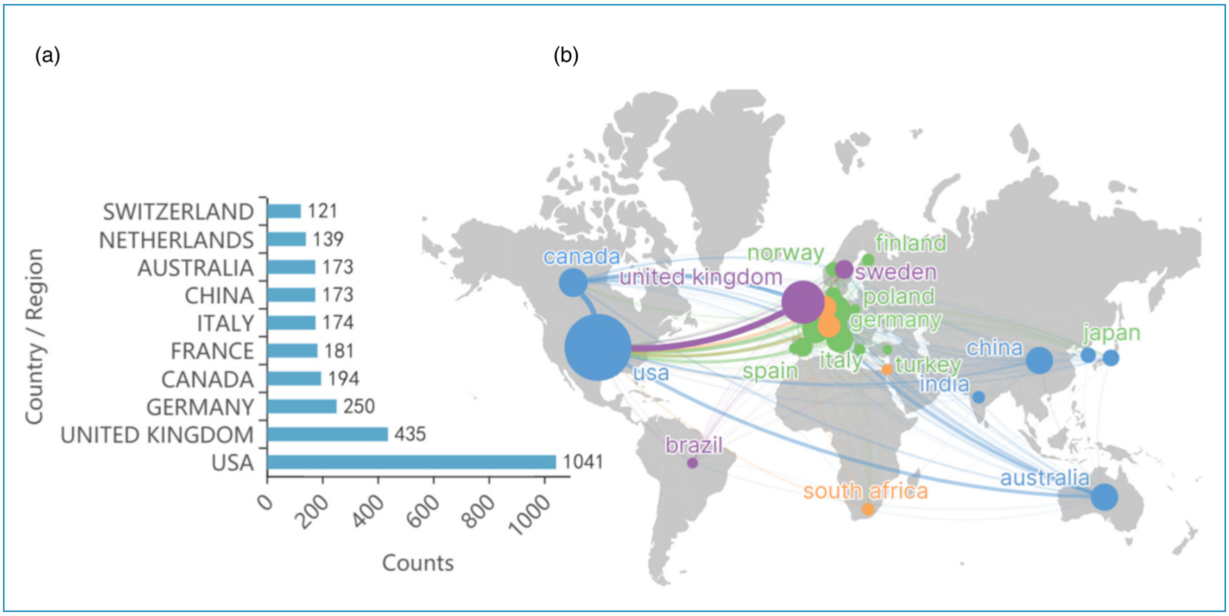


Figure 3. (a) Top 10 countries in publication counts. (b) Country collaboration network (a node represents a country, the links between nodes represent their collaboration relationships, different colors of nodes and links represent different research clusters).

more clearly. In Figure 3b, the color of the nodes corresponds to the countries/regions they represent, the size of the nodes indicates the volume of publications, and the darkness and thickness of the lines represent the strength of the collaboration, with the same color signifying a cluster. The most frequent collaborators were the UK (frequency = 938), the USA (frequency = 917), and Germany (frequency = 632). Notable strong collaborative relationships were observed between the USA and the UK, the USA and Canada, the UK and Germany, and the USA and Germany. Representative clustered collaboration networks include: (1) the USA, Canada, China, Australia, India, etc.; (2) the UK, Brazil, and Sweden, etc.; (3) Germany, France, Italy, Belgium, Denmark, Norway, etc. In the research on data

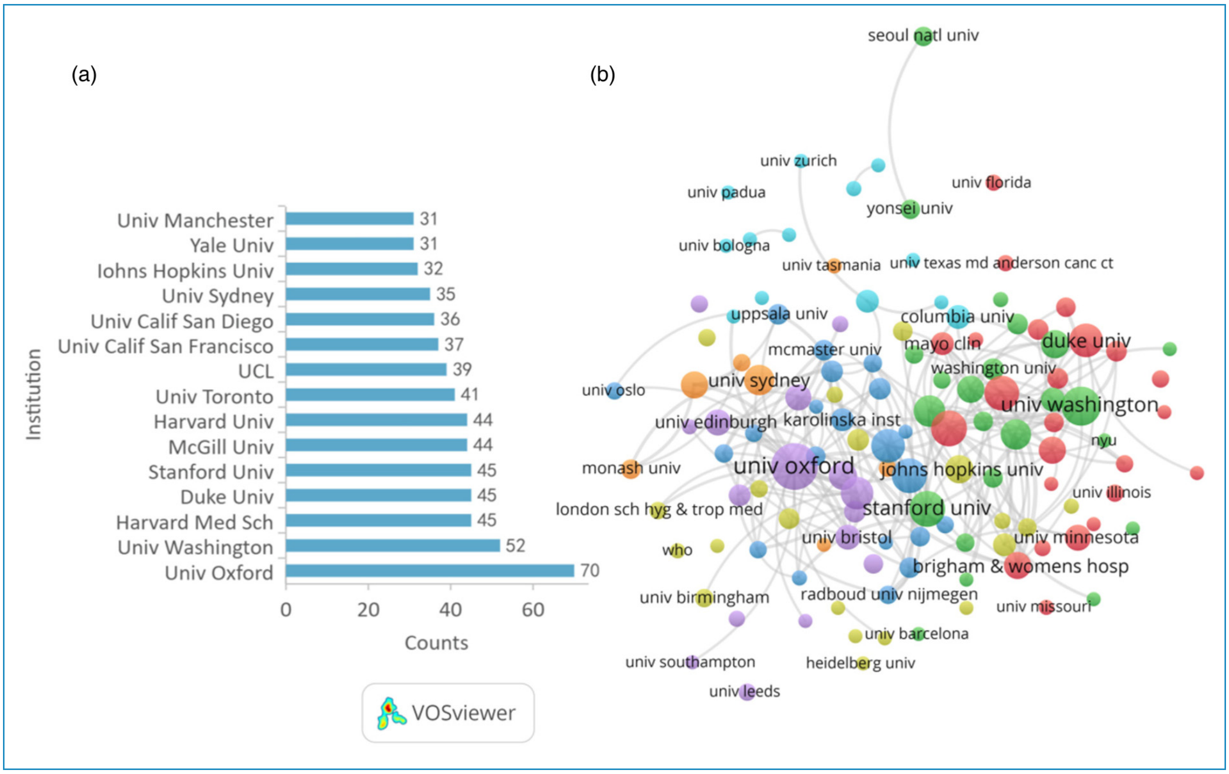


Figure 4. (a) Top 15 institutions in publication counts. (b) Institution collaboration network (a node represents an institution, the links between nodes represent their collaboration relationships, different colors of nodes represent different research clusters).

governance and open sharing in the fields of life sciences and medicine, the USA, the UK, and Germany have made significant contributions and have engaged in frequent collaborations.

Analysis of institution publications

A visualization analysis of the institutions involved in publications revealed that between 2000 and 2024, a total of 4629 institutions participated in the research on data governance and open sharing in the fields of life sciences and medicine. Figure 4 displays the top 15 institutions by publication volume and the institutional collaboration network. The University of Oxford in the UK had the highest publication volume, followed by the University of Washington, Harvard Medical School, Duke University, and Stanford University in the USA. There were 124 institutions with 10 or more publications, which formed seven clusters. The main clusters included: (1) a red cluster predominantly consisting of Harvard University, Harvard Medical School, and Duke University from the USA; (2) a green cluster predominantly consisting of the University of Washington, University of California, San Francisco, and University of California, San Diego from the USA; (3) a blue cluster predominantly consisting of the University of Toronto and McGill University from Canada; (4) a yellow cluster predominantly consisting of Johns Hopkins University from the USA; (5) a purple cluster predominantly consisting of the University of Oxford, University College London, and the University of Manchester from the UK. It is evident that in the research on data governance and open sharing in the fields of life sciences and medicine, research groups led by universities from the USA, the UK, and Canada were highly engaged and frequently collaborated.

Analysis of author publications

A statistical analysis of author publication records revealed that between 2000 and 2024, a total of 14,484 authors participated in the publication of articles on data governance and open sharing in the fields of life sciences and medicine. Figure 5a presents the top 10 authors by publication volume, with Amy L. McGuire leading the list with 14 publications, followed by Lucila Ohno-Machado with 13 publications. There were 201 authors with three or more publications, who formed 57 clusters. Figure 5b illustrates the collaboration and clustering among these 201 authors. In the figure, the size of the circles represents the number of publications, the darkness of the lines indicates the strength of collaboration, and the

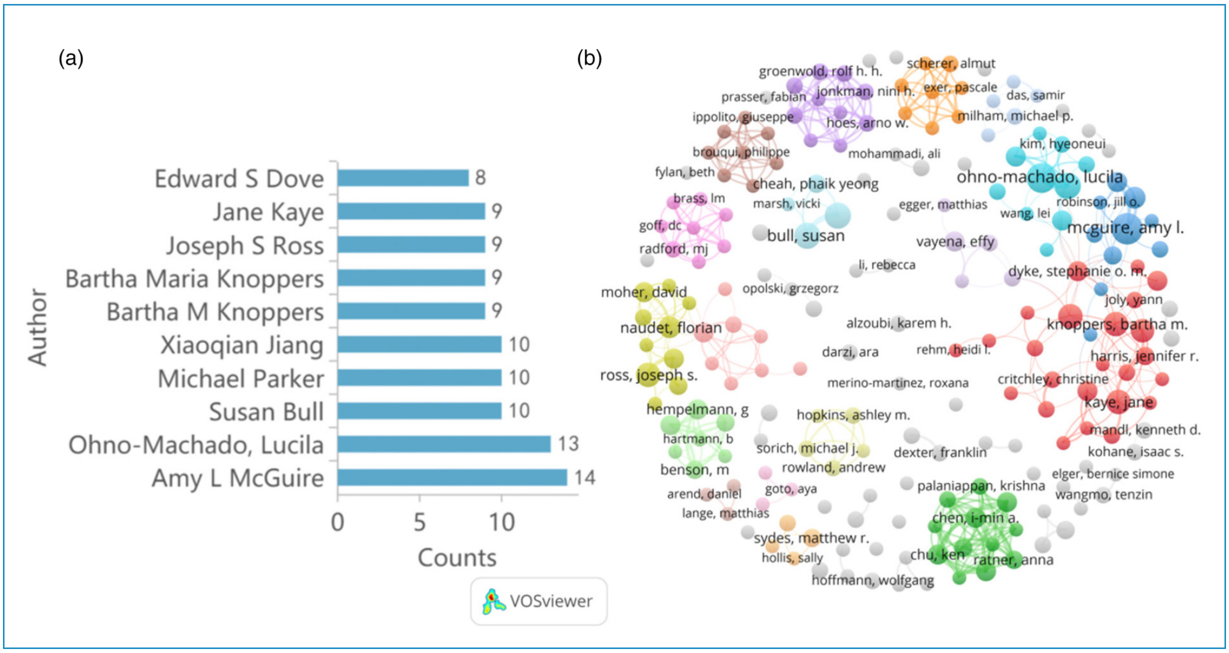


Figure 5. (a) Top 10 authors in publication counts. (b) Author collaboration network (a node represents an author, the links between nodes represent their collaboration relationships, different colors of nodes represent different research clusters).

colors represent different clusters. It is observable that research on data governance and open sharing in the fields of life sciences and medicine is often conducted in the form of small teams or individual work, with only a few teams showing intersecting collaborations. The majority of the work exists in the form of independent research by individual teams. This suggests that there is a need to enhance collaboration among individuals and between teams within the field, in pursuit of broader and larger-scale research partnerships.

Analysis of journal publications

This study included literature from 1022 journals. Table 1 lists the top 10 journals by publication volume, along with their publication counts, countries of origin, five-year impact factors, and JCR categories. These publishers are predominantly from the UK and the USA, with all of the journals located in the Q1 and Q2 quartiles of the Journal Citation Reports. The most frequently published journals in the field of data governance and open sharing in life sciences and medicine are the *Journal of Medical Internet Research* and the *Journal of the American Medical Informatics Association*. The journal with the highest impact factor is the *Journal of Medical Internet Research*. They are both positioned within the first quartile of the Journal Citation Reports. Researchers can identify potential journals for submission by referring to the most frequently published journals in the field of data governance and open sharing research in life sciences and medicine.

Analysis of references

Analysis of most cited references. The 2529 articles included in this study cited a total of 71,842 references. Table 2 presents the top 10 references by citation frequency. The most frequently cited document was the Comment published in Scientific Data titled 'The FAIR Guiding Principles for Scientific Data Management and Stewardship'.[15] This Comment noted that a workshop named 'Jointly Designing a Data Fairport' was held in Leiden, the Netherlands, in 2014, where experts collaboratively drafted the FAIR principles. After refinement and improvement by the FAIR Working Group, this Comment officially released the FAIR principles for the first time. Among the remaining papers, two addressed the identification of individuals through genetic information,[16,17] three collected the attitudes of the public or participants toward personal data sharing,[18–20] one discussed the attitudes of scientists toward data sharing,[21] and the themes of the remaining three papers were as follows: discussing how to share data more responsibly,[22] advocating for data sharing to improve public health,[23] and GlaxoSmithKline's Clinical Trials decision to share clinical trial data.[24]

Analysis of citation bursts. Bibliographic burst refers to a publication's citation count that is significantly higher than usual for a duration of at least two years, which can be used to explore emerging hotspots and research frontiers in a field of study.[25] The blue line in Figure 6 represents the observation period from 1990 to 2022, while the red line indicates the burst time of cited documents. The

Table 1. Top 10 journals in the field of data governance and open sharing in life sciences and medicine.

Rank	Source	Article	Country	IF	JCR
1	*Journal of Medical Internet Research*	42	Canada	6.7	Q1
2	*Journal of the American Medical Informatics Association*	42	USA	5.8	Q1
3	*BMC Bioinformatics*	40	UK	3.6	Q2
4	*BMJ Open*	40	UK	2.7	Q1
5	*Bioinformatics*	40	UK	7.6	Q1
6	*International Journal of Environmental Research and Public Health*	31	Switzerland	4.8	Q1
7	*International Journal of Medical Informatics*	26	Netherlands	4.6	Q1
8	*Computational Intelligence and Neuroscience*	25	USA	3.9	Q2
9	*BMC Medical Informatics and Decision Making*	22	UK	3.9	Q2
10	*Journal of Biomedical Informatics*	21	USA	7.4	Q2

Note. IF: Five year impact factor; JCR: Journal Citation Reports.

publication with the highest burst value between 2000 and 2024 was ‘Amounts of DNA to Highly Complex Mixtures Using High-Density SNP Genotyping Microarrays’, published in *PLOS Genetics*, with a burst intensity of 10.03. Its citation frequency was markedly higher between 2010 and 2014 than before. The paper pointed out that even when the data shared in population genetics research are aggregate summary statistics (such as allele frequencies or genotype counts), these data still cannot completely obscure individual identities.[16] Additionally, the citation bursts of three papers continued until 2024, covering topics such as public attitudes toward the sharing of personal health data for research purposes,[18] the FAIR principles in data governance,[15] and global public perceptions of genomic data sharing.[26] ‘Public Responses to the Sharing and Linkage of Health Data for Research Purposes: A Systematic Review and Thematic Synthesis of Qualitative Studies’[18] found that there is widespread public support for health data sharing, but this support is conditional, depending on data confidentiality, the realization of public interest, informed consent mechanisms for data use, and trust and transparency in research organizations. ‘Global Public Perceptions of Genomic Data Sharing: What Shapes the Willingness to Donate DNA and Health Data?’[26] indicated that the global public’s willingness to donate their DNA and health data for research is relatively low. Familiarity with genetics and trust in data users were significantly associated with willingness to donate. It is evident that the FAIR principles in data governance and public attitudes toward the sharing of genetic and other health data may become potential frontiers in the field of data governance and open sharing in life sciences and medicine.

Analysis of keywords

Frequency analysis of keywords. Keywords were extracted and their frequencies were counted for literature on data governance and open sharing in the fields of life sciences and medicine published between 2000 and 2024. Figure 7 lists the top 10 keywords by frequency. Among these keywords, ‘data sharing’ appeared 261 times, ranking first, ‘privacy’ ranked third, ‘data management’ ranked sixth, ‘information’ and ‘quality’ ranked ninth and tenth, respectively. ‘Care’, ‘mortality’, ‘risk’, ‘outcomes’ and ‘health’ were important characteristic terms in the fields of life sciences and medicine. It is evident that in the context of data governance and open sharing, privacy has received the highest level of attention, and researchers are also very concerned about the quality of data.

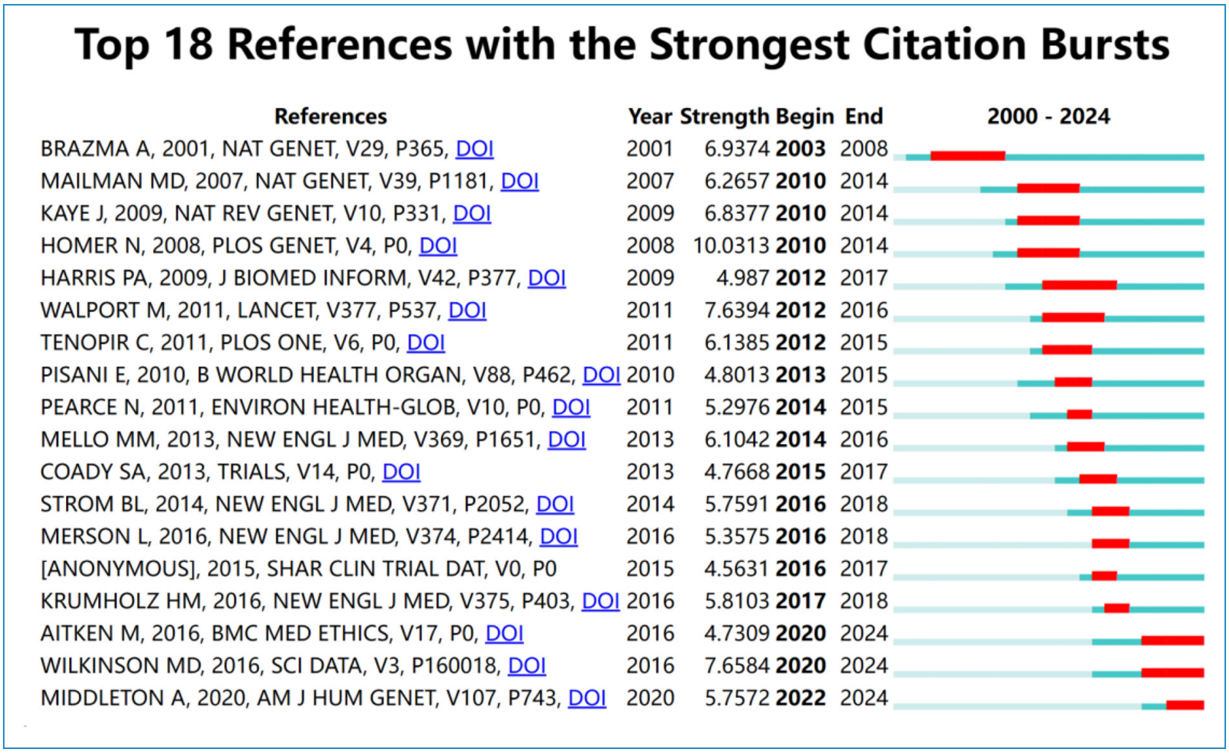
Co-occurrence analysis of keywords. Figure 8 presents a co-occurrence network constructed from high-frequency keywords, which can predominantly be categorized into three segments: the blue segment focused on ‘data sharing’, the green segment focused on ‘data management’, and the purple segment centered around characteristic terms of the life sciences and medicine fields such as ‘care’ and ‘health’. Within the data sharing module, keywords with high co-occurrence intensity included ‘privacy’, ‘ethics’, ‘consent’, ‘attitudes’ and ‘trust’. This indicates that privacy protection, informed consent, public attitudes, and

Table 2. Top 10 cited publications in the field of data governance and open sharing in life sciences and medicine.

Rank	Title	Year	Journal	First author	Cited counts
1	Comment: The FAIR Guiding Principles for Scientific Data Management and Stewardship	2016	*Scientific Data*	Wilkinson, Mark D.	64
2	Preparing for Responsible Sharing of Clinical Trial Data	2013	*New England Journal of Medicine*	Mello, Michelle M.	25
3	Public Responses to the Sharing and Linkage of Health Data for Research Purposes: A Systematic Review and Thematic Synthesis of Qualitative Studies	2016	*BMC Medical Ethics*	Aitken, Mhairi	21
4	Clinical Trial Participants' Views of the Risks and Benefits of Data Sharing	2018	*New England Journal of Medicine*	Mello, Michelle M.	21
5	Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures Using High-Density SNP Genotyping Microarrays	2008	*PLOS Genetics*	Homer, Nils	20
6	Identifying Personal Genomes by Surname Inference	2013	*Science*	Gymrek, Melissa	20
7	Sharing Research Data to Improve Public Health	2011	*Lancet*	Walport, Mark	19
8	Data Sharing by Scientists: Practices and Perceptions	2011	*PLoS ONE*	Tenopir, Carol	19
9	A Systematic Literature Review of Individuals' Perspectives on Broad Consent and Data Sharing in the USA	2016	*Genetics in Medicine*	Garrison, Nanibaa' A.	19
10	Access to Patient-Level Data from GlaxoSmithKline Clinical Trials	2013	*New England Journal of Medicine*	Nisen, Perry	18

trust toward data sharing, along with other ethical issues, are key factors affecting data sharing and are among the most closely watched topics by researchers in this area. In the data management module, keywords with high co-occurrence intensity included 'big data', 'database', 'machine learning', 'cancer' and 'standards'. In the fields of life sciences and medicine, there exists a vast amount of big data that is distinguished by its volume, velocity, variety, and variability. Discovering and extracting valuable scientific data from big data, utilizing databases for storage, management, and application, and applying technologies such as machine learning for data analysis are key to data governance. The high correlation between the keywords of 'cancer' and data governance modules suggests that cancer may be a disease of particular concern in the context of data governance. Additionally, ensuring that data conforms to standard specifications is also a hot topic in data governance. In the life sciences and medical modules, keywords with high co-occurrence intensity included 'care', 'health', 'outcomes', 'mortality' and 'risk', indicating that data governance and open sharing in the field of life sciences and medicine primarily serve the health and well-being of individuals.

Analysis of the research themes evolution. Keyword time zone maps reflect the evolutionary trends of research topics over time, revealing the updating of knowledge within research fields and their mutual influences.[27] The keyword time zone map, as shown in Figure 9, was divided into two-year intervals, where the size of the nodes represented the frequency of keyword co-occurrence, and the lines represented the relationships of co-occurrence between keywords. In the field of life sciences and medicine, the themes of focus from 2000 to 2009 included 'therapy', 'mortality', 'prevalence', 'care', 'disease' and 'outcome', which primarily emphasized passive treatment approaches to outcomes such as illness and death. From 2010 to 2019, the themes of focus shifted to 'health', 'diagnosis', 'epidemiology', 'cancer', 'electronic health record' and 'genomics', indicating a proactive pursuit of health



Top 18 References with the Strongest Citation Bursts

References	Year	Strength	Begin	End	2000 - 2024
BRAZMA A, 2001, NAT GENET, V29, P365, DOI	2001	6.9374	**2003**	2008	
MAILMAN MD, 2007, NAT GENET, V39, P1181, DOI	2007	6.2657	**2010**	2014	
KAYE J, 2009, NAT REV GENET, V10, P331, DOI	2009	6.8377	**2010**	2014	
HOMER N, 2008, PLOS GENET, V4, P0, DOI	2008	10.0313	**2010**	2014	
HARRIS PA, 2009, J BIOMED INFORM, V42, P377, DOI	2009	4.987	**2012**	2017	
WALPORT M, 2011, LANCET, V377, P537, DOI	2011	7.6394	**2012**	2016	
TENOPIR C, 2011, PLOS ONE, V6, P0, DOI	2011	6.1385	**2012**	2015	
PISANI E, 2010, B WORLD HEALTH ORGAN, V88, P462, DOI	2010	4.8013	**2013**	2015	
PEARCE N, 2011, ENVIRON HEALTH-GLOB, V10, P0, DOI	2011	5.2976	**2014**	2015	
MELLO MM, 2013, NEW ENGL J MED, V369, P1651, DOI	2013	6.1042	**2014**	2016	
COADY SA, 2013, TRIALS, V14, P0, DOI	2013	4.7668	**2015**	2017	
STROM BL, 2014, NEW ENGL J MED, V371, P2052, DOI	2014	5.7591	**2016**	2018	
MERSON L, 2016, NEW ENGL J MED, V374, P2414, DOI	2016	5.3575	**2016**	2018	
[ANONYMOUS], 2015, SHAR CLIN TRIAL DAT, V0, P0	2015	4.5631	**2016**	2017	
KRUMHOLZ HM, 2016, NEW ENGL J MED, V375, P403, DOI	2016	5.8103	**2017**	2018	
AITKEN M, 2016, BMC MED ETHICS, V17, P0, DOI	2016	4.7309	**2020**	2024	
WILKINSON MD, 2016, SCI DATA, V3, P160018, DOI	2016	7.6584	**2020**	2024	
MIDDLETON A, 2020, AM J HUM GENET, V107, P743, DOI	2020	5.7572	**2022**	2024	

Figure 6. Top 18 references with the strongest citation bursts (blue line represents the time intervals when cited literature appears, red bars indicate that the number of citations of the literature suddenly increased during that period).

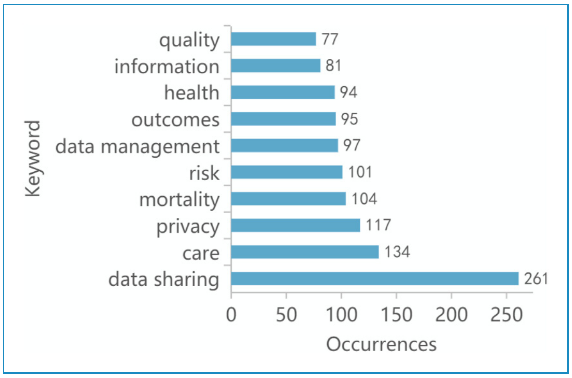


Figure 7. Top 10 most frequent keywords.

from the entry points of chronic diseases, electronic health records, and genomics. Since 2020, the themes of ‘public health’ and ‘COVID-19’ have come to the forefront, mainly influenced by the COVID-19, which has led to increased attention on public health issues such as pandemics. On the research content level of data governance and open sharing, the primary focus from 2000 to 2007 was on ‘database’, ‘data management system’, and ‘bioinformatics’ related to data management. From 2008 to 2021, the focus shifted toward content related to data sharing, including ‘privacy’, ‘ethics’, ‘consent’, ‘attitude’, ‘trust’, and ‘safety’. Since 2022, the primary focus has been on content related to data quality and data governance. On the technological level of data governance and open sharing, ‘big data’ emerged in 2014, ‘blockchain’ in 2018, ‘artificial intelligence’ in 2020, and ‘machine learning’ in 2022.

Clustering analysis of keywords. Importing the literature records into the CiteSpace software, eight clusters were formed, as depicted in Figure 10. These clusters are ‘ethics’, ‘data management’, ‘outliers’, ‘data sharing’, ‘clinical trial’, ‘prevalence’, ‘biopsy’ and ‘acute coronary syndrome’. The ‘ethics’ cluster indicates that the research on data governance and open sharing in life sciences and medicine is primarily concerned with related ethical issues. The ‘outliers’ cluster suggests that outliers in the dataset may affect data quality and could potentially aid in identity recognition, leading to the disclosure of personal privacy. The emergence of the ‘clinical trial’, ‘prevalence’, ‘biopsy’ and ‘acute coronary syndrome’ clusters indicates that data from clinical trials, epidemiological diseases, cancer and tumor, and cardiovascular diseases may receive higher attention in data governance and open sharing due to their large volume, importance, and particularity.

Citation burst analysis of keywords. Figure 11 presents the top 19 keywords with the highest burst intensity from

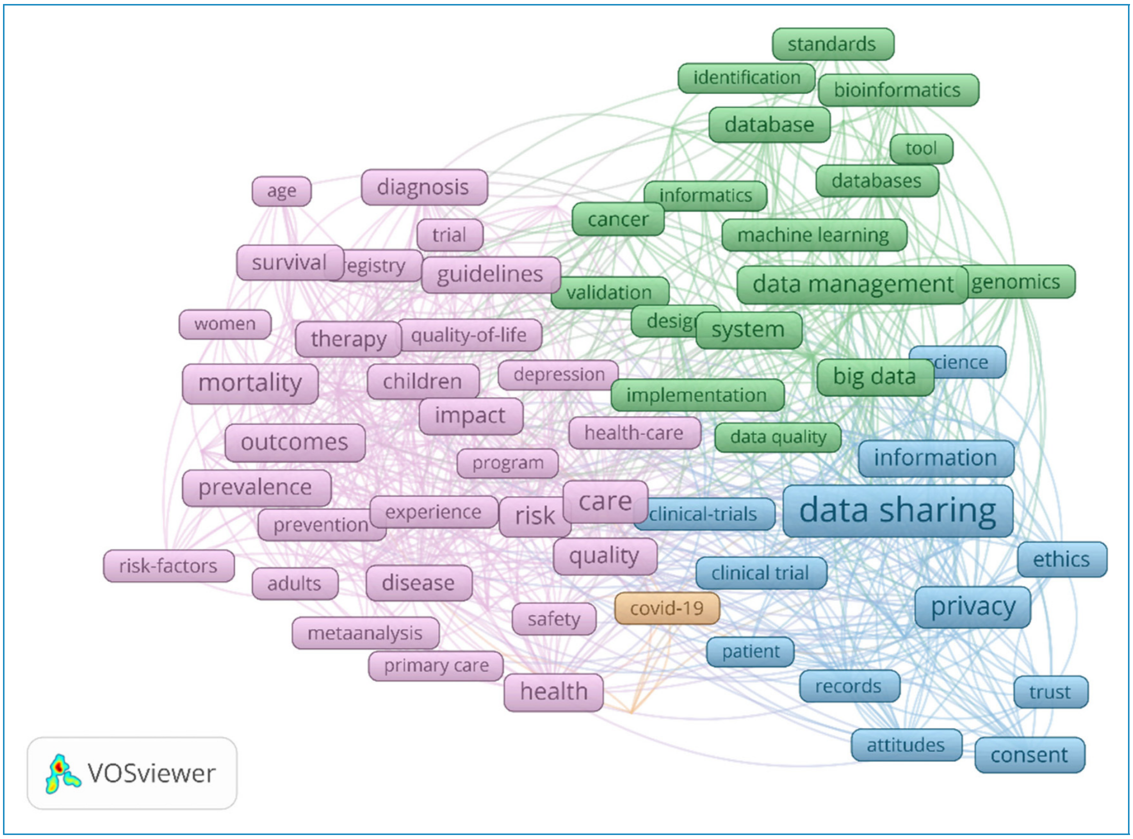


Figure 8. Keywords co-occurrence network (a node represents a keyword, the links between nodes represent the co-occurrence relationships between keywords, and the different colors of nodes and links indicate different thematic clusters).

2000 to 2024. The keyword with the highest burst value was ‘COVID-19’, which emerged in 2021 and continued to the present, indicating that the sharing and governance of data related to novel coronavirus is significant to mitigate the spread of the COVID-19 pandemic. ‘Data management’ emerged between 2001 and 2004, while ‘data governance’ burst from 2022 to 2024. Data management focuses on the technical processing and operation of data, whereas data governance emphasizes strategic management and oversight. The evolution from data management to data governance reflects a shift in data’s role from an initial tool for recording and storage to a critical asset and strategic resource in scientific research and social development. The evolution of burst terms across different time periods reveals that the construction of databases was a research focus during 2002–2010. From 2005 to 2011, researchers began to pay attention to the design and optimization of data management systems. Between 2015 and 2017, the sharing and management of clinical trial data became a research hotspot. During 2016–2019, the sharing and management of data related to quality of life received widespread attention. Data security issues were a significant concern from 2018 to 2022, with the application of blockchain and artificial intelligence in data sharing and governance becoming a hot topic. Keywords that burst until 2024 included ‘epidemiology’, ‘adult’, ‘attitude’, ‘public health’, ‘obesity’, ‘trust’, ‘covid-19’, ‘model’, ‘health’ and ‘data governance’. This suggests that epidemiology and public health, public trust and attitudes toward data sharing, sharing and governance of data related to chronic diseases such as diabetes, and the construction of data governance models may become future research hotspots and frontiers.

Discussion

Overview of publications in the research area

During the period from 2000 to 2024, the annual number of publications on data governance and open sharing in the fields of life sciences and medicine has shown an overall trend of fluctuating growth. The number of publications surpassed 100 articles in 2014 and reached a peak in 2021 before experiencing a slight decline. The overall increasing trend in publication volume can be attributed to several potential reasons: (1) With the rapid development of omics technologies and information technology, the amount of data in life sciences and medicine has undergone an explosive growth, leading to a profound transformation toward a data-intensive fourth scientific paradigm. (2) The application of big data and artificial

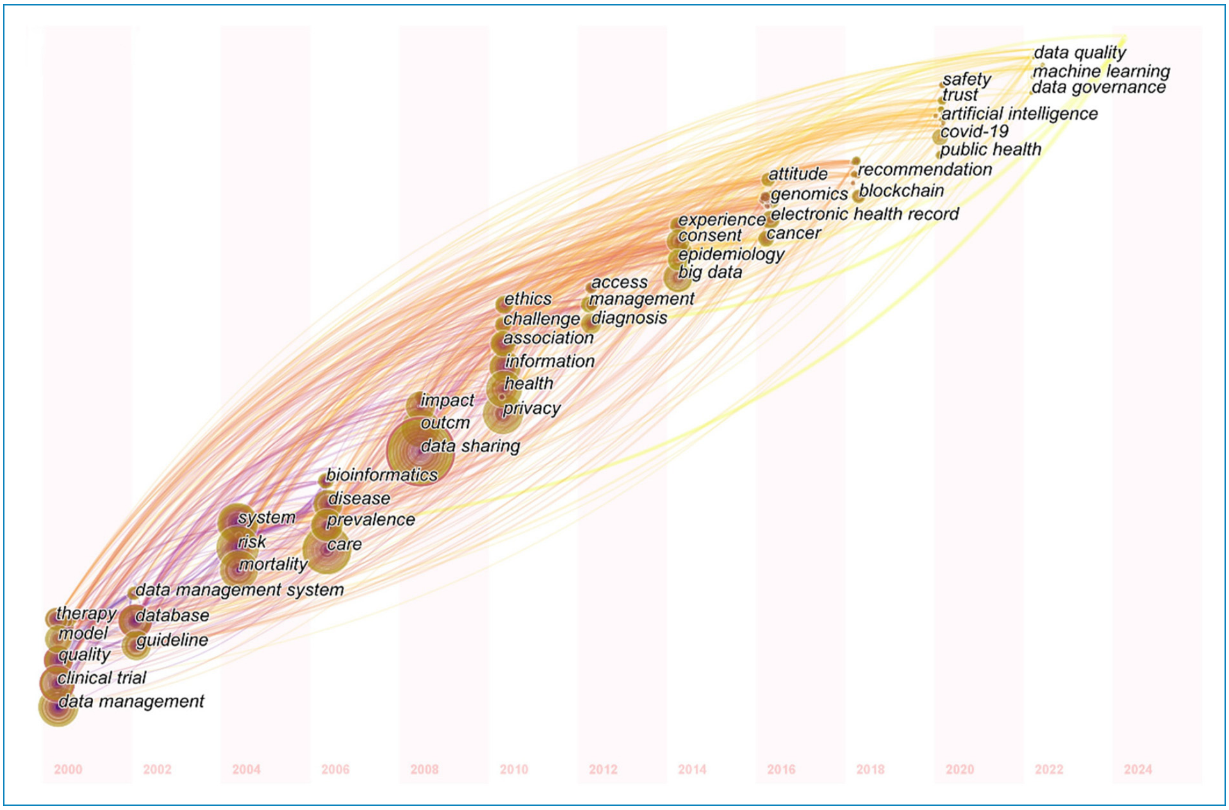


Figure 9. Time zone view of keywords (taking two years as a time slice, the nodes within the slice represent the high-frequency keywords of that period, and the links between nodes represent the co-occurrence relationships between keywords).

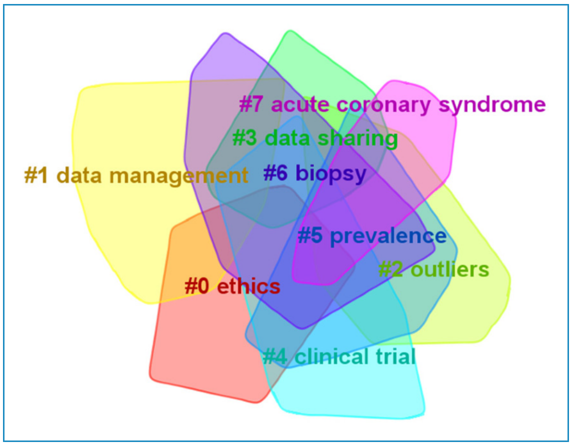
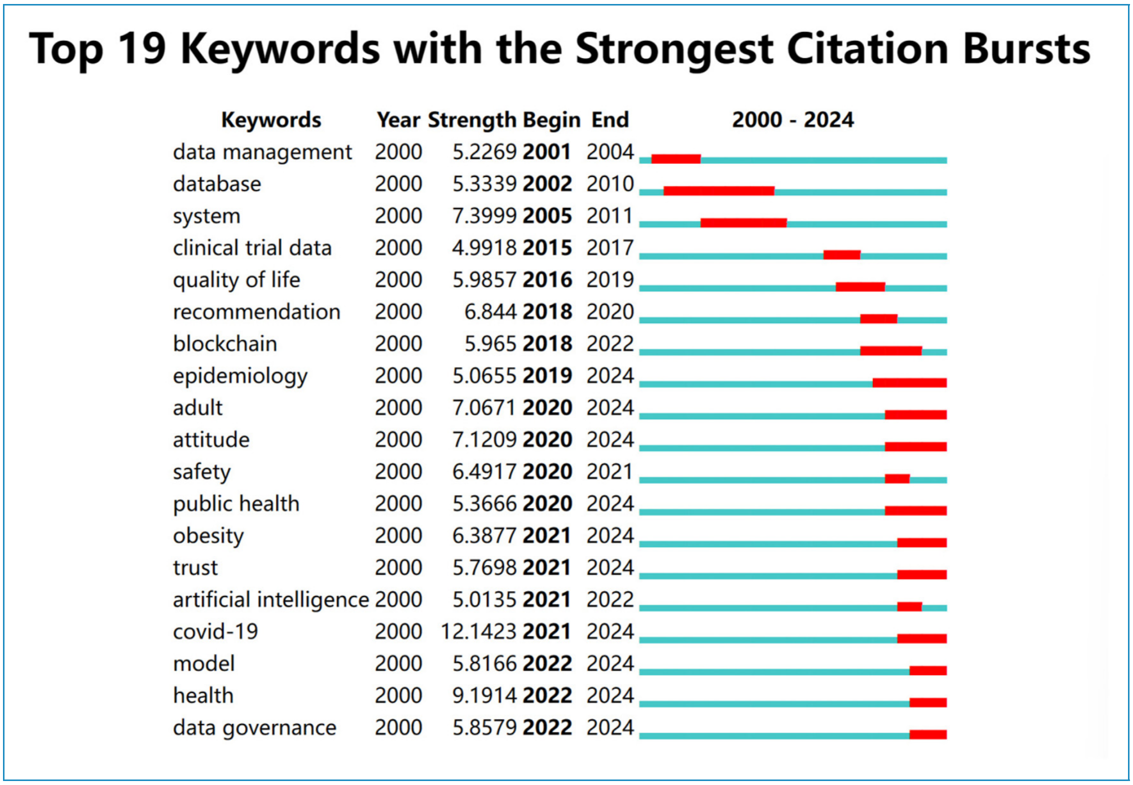


Figure 10. Keywords clustering map (each color block represents a thematic cluster generated by keywords).

intelligence technologies has made the analysis and processing of massive data possible and convenient,[28] thereby promoting research on data governance and open sharing. (3) The rise of the open science movement has facilitated data sharing, with an increasing number of datasets becoming openly accessible, enhancing data transparency and availability.[29] (4) The growing attention to health and the increasing emphasis on life sciences and medical research worldwide have led to greater investment of resources by governments and research institutions, which has in turn stimulated research and publication in these areas. The annual number of published papers fell slightly after reaching a peak in 2021, which may be due to the transfer of research hotspots and fluctuations in research funds. It is also possible that some issues have impacted the progress and publication of research in the field, including data privacy protection, data classification and standardization, human genetic resource data sharing risk, and so on. Additionally, the trend in annual publication volume correlates highly with the exponential growth function prediction model, suggesting that there is a certain regularity and predictability to the growth of research in this field. This strong correlation may imply that as time goes on, the importance of data governance and open sharing will be further recognized and accepted, and the open sharing and scientific governance of data is an irreversible trend of the times.

In the context of national publication output, the USA, the UK, Germany and Canada are the leading countries in terms of publication volume and are also the most frequent collaborators with other nations. Eighty percent of the top 10 countries by publication volume are located in the



Top 19 Keywords with the Strongest Citation Bursts

Keywords	Year	Strength	Begin	End	2000 - 2024
data management	2000	5.2269	**2001**	2004	
database	2000	5.3339	**2002**	2010	
system	2000	7.3999	**2005**	2011	
clinical trial data	2000	4.9918	**2015**	2017	
quality of life	2000	5.9857	**2016**	2019	
recommendation	2000	6.844	**2018**	2020	
blockchain	2000	5.965	**2018**	2022	
epidemiology	2000	5.0655	**2019**	2024	
adult	2000	7.0671	**2020**	2024	
attitude	2000	7.1209	**2020**	2024	
safety	2000	6.4917	**2020**	2021	
public health	2000	5.3666	**2020**	2024	
obesity	2000	6.3877	**2021**	2024	
trust	2000	5.7698	**2021**	2024	
artificial intelligence	2000	5.0135	**2021**	2022	
covid-19	2000	12.1423	**2021**	2024	
model	2000	5.8166	**2022**	2024	
health	2000	9.1914	**2022**	2024	
data governance	2000	5.8579	**2022**	2024	

Figure 11. Top 19 keywords with the strongest citation bursts (blue line represents the time intervals when keywords appear, and red bars indicate the periods of a surge in citation volume).

Americas and Europe. Among the top 15 institutions in terms of publication output, nine are based in the USA, three in the UK, two in Canada, and one in Australia. Large-scale institutional clusters are primarily led by universities in the USA, the UK and Canada. Although Germany ranks third in publication output, none of its research institutions make it into the top 15 in terms of publication volume, nor has it formed a large-scale cluster dominated by German institutions. In terms of author publication output, four out of the top 10 authors are from the USA, three from the UK, two from Canada, and one from Australia, which is similar to the distribution of institutions. It is evident that in the field of data governance and open sharing in life sciences and medicine, the USA holds a leading position, with other high-contributing countries mainly located in the Americas and Europe. This may be related to their higher levels of economic development and investment in healthcare. Additionally, these countries have established a number of comprehensive data centers to facilitate the open sharing of life sciences and medical data, such as the NCBI, the EBI, and the Swiss Institute of Bioinformatics (SIB). They have developed relatively mature management systems and have fostered a robust data ecosystem.

The collaboration between countries shows a pattern of ‘transcontinental global linkage’ and ‘intracontinental regional clustering’. The former is exemplified by collaborations between the USA and the UK, the USA and China, and the USA and Australia. The latter is exemplified by collaborations between the USA and Canada, as well as within Western European countries. It can be observed that research in data governance and open sharing in the field of life sciences and medicine is primarily concentrated in economically developed countries. However, many developing countries, facing harsh living conditions and health issues, also have a demand for improving medical and health levels through data governance and open sharing. However, many developing countries face poor data governance and sharing conditions due to the lack of a universal data sharing platform or framework, the absence of guidelines or policies for data security and privacy protection, and the lack of awareness among researchers about the necessity of extensive data sharing.[30] It is recommended that developed countries enhance their radiating and leading role in this field, actively cooperate with institutions and scientists in developing countries, raise their awareness of data governance and sharing, support developing countries in participating in cross-border data flows, provide capacity building and technical assistance, and include the perspectives of developing countries in the formulation of global data governance and sharing rules, in order to reduce the digital divide and health inequalities, and promote a broader opening and sharing of data.

Analysis of the thematic evolution in the research area

The evolution of research themes in the field of life sciences and medical data governance and open sharing can be dissected from three perspectives. Firstly, examining the content of research within the life sciences and medical field, the focus of scholars has shifted from a disease-centric approach during 2000–2009 to a health-centric approach during 2010–2019. Since 2020, the emphasis has shifted toward epidemiology and public health issues. This reflects a paradigm change in the field from a 'passive treatment' to a 'proactive health' mindset, and from an 'individual' to a 'population' perspective. Governments, hospitals, and research institutions are urged to effectively manage and share resident health-related data, including physical examination data, chronic disease data, and nutritional data, to address scientific research demands centered on health management and promotion, as well as industrial needs for the development of health monitoring and management products. Concurrently, governments, hospitals, disease prevention and control agencies, and emergency management organizations should manage and share data on epidemics and public health, such as data related to the COVID-19 pandemic, to provide experiential references for potential future outbreaks and to deploy more proactive prevention and control measures.

From the perspective of research content in data governance and open sharing, the period from 2000 to 2007 marked a phase of prosperity for data management research. The years from 2008 to 2021 were characterized by a flourishing phase of data sharing research, and the period from 2022 to the present has seen the emergence of data governance. Data management involves the management of activities throughout the data lifecycle, while data governance encompasses the planning, decision-making, supervision, and control of data management.[31] The evolution from data management to data governance reflects the scholarly community's emphasis on data quality, security, and legal compliance, as well as their aspiration to create a sustainable data ecosystem. However, it is noteworthy that data management and data governance are closely interrelated and indispensable to each other. Although research focuses may vary across different periods, it is essential to ensure the rational coexistence of both. In order to achieve the goals of ensuring data quality and security, maximizing the value of data assets, and conducting data sharing and dissemination in a legal and compliant manner. Databases and data centers, in the process of storing, sharing, processing, and utilizing data, should not only manage the entire lifecycle of data at the micro-level but also adopt a more macro-perspective to facilitate the participation of diverse stakeholders in data governance. Additionally, they should integrate and utilize a variety of tools to enhance the quality, security, compliance, and ethicality of data.

In the realm of technological applications for data governance and open sharing, the year 2014 marked a significant turning point as big data and artificial intelligence dramatically entered the purview of scholars. Big data analytics can integrate diverse types of information, transforming vast amounts of data into actionable knowledge that aids in precision medicine, disease diagnosis, and risk warning.[32] Blockchain technology offers a potential decentralized distributed network for data sharing and governance,[33] yet it comes with inherent risks such as standards and interoperability issues, information privacy, and security concerns. Artificial intelligence possesses the capability to rapidly process large volumes of data as well as identify patterns and trends that may elude immediate human detection, thereby bringing additional possibilities to the governance and open sharing of data in the life sciences and medicine. Data sharing can also facilitate the collection of extensive data needed to train powerful and highly predictive AI models. However, the unique requirements for privacy and security in this domain impose certain limitations on data access, which to some extent hinders the development of robust AI tools.[34] Researchers in the fields of life sciences and medicine should keep abreast of and educate themselves on big data and artificial intelligence technologies. They should apply these technologies judiciously in the processing and utilization of data, ensuring that large volumes of data can serve the health needs of residents and contribute to societal well-being. Concurrently, they must be vigilant against the leakage and misuse of personal health-related data, paying close attention to the protection of data security and the privacy of residents.

Analysis of hot spots and frontiers in the research area

Analysis of highly cited publications, high-frequency keywords, keyword co-occurrences and clustering reveals that the FAIR principles, ethical issues such as informed consent and privacy protection, public attitudes toward data sharing, data quality, databases, and big data analytics are hot topics in the field of data governance and open sharing within life sciences and medicine. The burst detection feature of CiteSpace can identify emerging research frontiers.[35] Combining the burst detection of cited publications and keywords with the evolution trends of themes, potential research frontiers in this field are identified to include: the FAIR principles, data quality, public trust and attitudes toward data sharing, the application of artificial intelligence technology in data sharing and governance, sharing and governance of epidemiological and public health data, sharing and governance of data on chronic diseases such as diabetes, and the construction of data governance models.

The FAIR principles have clarified the objectives of scientific data management, and have gained widespread recognition from international stakeholders since their publication, marking a milestone in the development of scientific data guidelines. While the FAIR principles are not a sufficient set of principles for responsible data sharing, they are necessary.[36] Standardized data sharing in accordance with the FAIR principles forms the foundation for the application of new data-driven artificial intelligence analytical techniques.[37] Implementing the FAIR principles is crucial for enhancing data quality and maximizing the value of data. However, addressing ethical issues is a prerequisite and foundation for extensive data sharing. Internationally, significant attention is given to ethical concerns involved in data sharing, such as privacy protection and informed consent. Internationally, several open-access databases, such as the NCBI, UK Biobank, the Global Initiative on Sharing All Influenza Data (GISAID), Medical Information Mart for Intensive Care (MIMIC) and TCGA, adhere to specific legal and ethical standards. The NCBI, TCGA, and MIMIC, funded by the NIH in the USA, comply with the Health Insurance Portability and Accountability Act (HIPAA) regulations, which mandate the de-identification of Protected Health Information (PHI) to safeguard individual privacy. NCBI, UK Biobank, and GISAID ensure that all data processing activities are in accordance with the GDPR requirements, adhering to the principle of data minimization during data collection and emphasizing the protection of data subjects' rights.

The public's attitude toward sharing their personal health data also significantly influences the process of data sharing. Data quality is another key factor affecting data sharing. Establishing an objective and systematic data quality management system is one of the core tasks of data governance. This system is essential to ensure the reliability of data, mitigate the risks associated with erroneous data, reduce the costs of data management, and enhance the utilization rate of data.[38] Big data and artificial intelligence technologies have demonstrated unique value in addressing some issues in data governance and open sharing. Some scholars propose methods such as federated learning and collaborative learning, which enable the collaborative training of machine learning models on distributed devices without disclosing sensitive data, thereby aiding in resolving data privacy and compliance issues.[39] Recently, model construction has become a research hotspot and frontier, with complex and diverse model-related studies. These include using data-driven models for anomaly prediction and maintenance of medical facilities[40]; employing statistical predictive models to monitor and evaluate pediatric cancer data[41]; and training AI models to help identify potential health risk factors and disease diagnostic targets, discover new drugs and vaccines, and develop personalized treatment plans.[39]

The hot topics and frontiers in the research domain provide scholars with directions for future studies. There should be increased focus on data quality issues, ethical considerations in data usage, and the application of big data and artificial intelligence technologies. Concurrently, there should be a strengthened governance and sharing of data in critical areas such as public health and chronic diseases. Ethical issues surrounding data encompass informed consent, privacy protection, and public trust. The application of big data and artificial intelligence technologies should particularly concentrate on the utilization of large-scale data models in disease detection, diagnosis, and clinical treatment.

Conclusion

Between 2000 and 2024, the number of studies on data governance and open sharing in the fields of life sciences and medicine has increased annually, indicating the growing importance of research in this area. The USA leads as the country with the most research output in this field. The University of Oxford is the institution with the highest publication volume, Amy L. McGuire is the most active author, and the *Journal of Medical Internet Research* and the *Journal of the American Medical Informatics Association* are the most frequent publication outlets. The most cited reference is 'Comment: The FAIR Guiding Principles for Scientific Data Management and Stewardship'. Topics such as the FAIR principles, ethical issues, public attitudes toward data sharing, data quality, databases, and big data analytics technologies are hot topics in this field. Potential research frontiers include the FAIR principles, data quality, public trust and attitudes toward data sharing, the application of artificial intelligence technology in data sharing and governance, sharing and governance of epidemiological and public health data, sharing and governance of data related to chronic diseases such as diabetes, and the construction of data governance models.

Acknowledgements: The authors want to thank CiteSpace and VOSviewer for free access by researchers.

Contributorship: Zhimin Hu contributed to conceptualization, funding acquisition, project administration, supervision and writing–review and editing. Yanrui Qiu contributed to conceptualization, data curation, methodology, software, visualization and writing–original draft. All authors contributed to the article and reviewed the submitted version.

Data availability: The data in this study is not sensitive and is accessible in the public domain. All the data used in the study have been included in the article.

Declaration of conflicting interests: The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Ethical approval: This article does not contain any studies with human participants or animals performed by any of the authors.

Funding: The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by the Noncommunicable Chronic Diseases-National Science and Technology Major Project (2023ZD0509701), and Medical and Health Technology Innovation Project of Chinese Academy of Medical Sciences (2021-I2M-1-057).

Guarantor: The guarantor of the study, Zhimin Hu, had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

ORCID iDs: Yanrui Qiu https://orcid.org/0009-0006-2314-4885
Zhimin Hu https://orcid.org/0000-0001-6294-219X

References

1. Liu L and Si L. Scientific data governance practices: content systems and development trend. *Info Stud Theory Appl* 2023; 46: 175–182.
2. Zhang GQ, Li YX, Wang ZF, et al. New challenges and trends in bio-med big data. *Subj Field* 2018; 33: 853–860.
3. Piasecki J and Cheah PY. Ownership of individual-level health data, data sharing, and data governance. *BMC Med Ethics* 2022; 23: 104. Epub 20221029.
4. Craig DW, Goor RM, Wang ZY, et al. Assessing and managing risk when sharing aggregate genetic variant data. *Nat Rev Genet* 2011; 12: 730–736.
5. Pham-Kanter G, Zinner DE and Campbell EG. Codifying collegiality: recent developments in data sharing policy in the life sciences. *PLoS ONE* 2014; 9: e108451.
6. Shabani M. The data governance act and the EU's move towards facilitating data sharing. *Mol Syst Biol* 2021; 17: e10229.
7. Lv JY, Li YM, Shi SQ, et al. Frontier and hotspot evolution in cardiorenal syndrome: a bibliometric analysis from 2003 to 2022. *Curr Probl Cardiol* 2023; 48: 101238.
8. Jiang ST, Liu YG, Zheng H, et al. Evolutionary patterns and research frontiers in neoadjuvant immunotherapy: a bibliometric analysis. *Int J Surg* 2023; 109: 2774–2783.
9. Zhang LL, Ling J and Lin MW. Carbon neutrality: a comprehensive bibliometric analysis. *Environ Sci Pollut Res* 2023; 30: 45498–45514.
10. Zhao WJ, Tan RZ, Gao J, et al. Research on the global trends of COVID-19 associated acute kidney injury: a bibliometric analysis. *Renal Fail* 2024; 46: 2338484.
11. Li XY, Su JM, Wang H, et al. Bibliometric analysis of artificial intelligence in wastewater treatment: current status, research progress, and future prospects. *J Environ Chem Eng* 2024; 12: 113152.
12. Keh CG, Gan PT, Gamal AAM, et al. Financial development-economic growth nexus: a bibliometric analysis. *Environ Dev Sustainability* 2024: 1–28.
13. Lee JY and Syn SY. Global research trends in research data management: a bibliometrics approach. *J Librariansh Inf Sci* 2024.
14. Pradhan P and Zala LN. Bibliometrics analysis and comparison of global research literatures on research data management extracted from Scopus and Web of Science during 2000–2019. *Libr Philos Pract* 2021.
15. Wilkinson MD, Dumontier M, Aalbersberg IJ, et al. Comment: the FAIR guiding principles for scientific data management and stewardship. *Sci Data* 2016; 3: 160018.
16. Homer N, Szelinger S, Redman M, et al. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genet* 2008; 4: e1000167. Epub 20080829.
17. Gymrek M, McGuire AL, Golan D, et al. Identifying personal genomes by surname inference. *Science* 2013; 339: 321–324.
18. Aitken M, Jorre JD, Pagliari C, et al. Public responses to the sharing and linkage of health data for research purposes: a systematic review and thematic synthesis of qualitative studies. *BMC Med Ethics* 2016; 17: 73.
19. Mello MM, Lieou V and Goodman SN. Clinical trial participants' views of the risks and benefits of data sharing. *N Engl J Med* 2018; 378: 2202–2211.
20. Garrison NA, Sathe NA, Antommaria AHM, et al. A systematic literature review of individuals' perspectives on broad consent and data sharing in the United States. *Genet Med* 2016; 18: 663–671.
21. Tenopir C, Allard S, Douglass K, et al. Data sharing by scientists: practices and perceptions. *PLoS ONE* 2011; 6: e21101.
22. Mello MM, Francer JK, Wilenzick M, et al. Preparing for responsible sharing of clinical trial data. *N Engl J Med* 2013; 369: 1651–1658.
23. Walport M and Brest P. Sharing research data to improve public health. *Lancet* 2011; 377: 537–539.
24. Nisen P and Rockhold F. Access to patient-level data from GlaxoSmithKline clinical trials. *N Engl J Med* 2013; 369: 475–478.
25. Qu M, Xu Y and Lu L. Global research evolution and frontier analysis of artificial intelligence in brain injury: a bibliometric analysis. *Brain Res Bull* 2024; 209: 110920. Epub 20240305.
26. Middleton A, Milne R, Almarri MA, et al. Global public perceptions of genomic data sharing: what shapes the willingness to donate DNA and health data?. *Am J Hum Genet* 2020; 107: 743–752.
27. Sun L, Wu L and Qi P. Global characteristics and trends of research on industrial structure and carbon emissions: a bibliometric analysis. *Environ Sci Pollut Res Int* 2020; 27: 44892–44905. Epub 20200929.
28. Mahmud M, Kaiser MS, Hussain A, et al. Applications of deep learning and reinforcement learning to biological data. *IEEE Trans Neural Netw Learn Syst* 2018; 29: 2063–2079.
29. Rockhold F, Bromley C, Wagner EK, et al. Open science: the open clinical trials data journey. *Clin Trials* 2019; 16: 539–546.

30. Kaewkungwal J, Adams P, Sattabongkot J, et al. Issues and challenges associated with data-sharing in LMICs: perspectives of researchers in Thailand. *Am J Trop Med Hyg* 2020; 103: 528–536.
31. Sheng XP and Song DC. A comparative analysis of data management and data governance and its enlightenment to the formulation of open sharing policies of scientific data. *Documentation Inf Knowledge* 2020; 64: 4–10.
32. Dong JC, Wu HQ, Zhou D, 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.
33. Shabani M. Blockchain-based platforms for genomic data sharing: a de-centralized approach in response to the governance problems?. *J Am Med Inform Assoc* 2019; 26: 76–80.
34. Pereira T, Morgado J, Silva F, et al. Sharing biomedical data: strengthening AI development in healthcare. *Healthcare* 2021; 9: 827.
35. Chen CM. Citespace II: detecting and visualizing emerging trends and transient patterns in scientific literature. *J Am Soc Inf Sci Technol* 2006; 57: 359–377.
36. Boeckhout M, Zielhuis GA and Bredenoord AL. The FAIR guiding principles for data stewardship: fair enough?. *Eur J Hum Genet* 2018; 26: 931–936. Epub 20180517.
37. Reer A, Wiebe A, Wang X, et al. FAIR human neuroscientific data sharing to advance AI driven research and applications: legal frameworks and missing metadata standards. *Front Genet* 2023; 14: 1086802. Epub 20230313.
38. Lee S, Roh GH, Kim JY, et al. Effective data quality management for electronic medical record data using SMART DATA. *Int J Med Inf* 2023; 180: 105262.
39. Tajabadi M, Grabenhenrich L, Ribeiro A, et al. Sharing data with shared benefits: artificial intelligence perspective. *J Med Internet Res* 2023; 25: 6.
40. Zhou HP, Liu QL, Liu HW, et al. Healthcare facilities management: a novel data-driven model for predictive maintenance of computed tomography equipment. *Artif Intell Med* 2024; 149: 102807.
41. Martínez-Salazar J and Toledano-Toledano F. Comparative analysis of three predictive models of performance indicators with results-based management: cancer data statistics in a National Institute of Health. *Cancers (Basel)* 2023; 15: 4649.