Learned Publishing

LEARNED PUBLISHING Open Access

ORIGINAL ARTICLE OPEN ACCESS

Data Availability Statements in Mega Journals: A Comparative Analysis of Global and Korea-Affiliated Publications in Health and Medical Research

Sanghee Oh | Yunseo Park | Seyun Sim

Department of Library and Information Science, Sungkyunkwan University, Seoul, South Korea

Correspondence: Sanghee Oh (sangheeoh@skku.edu)

Received: 20 January 2026 | **Revised:** 19 May 2026 | **Accepted:** 12 June 2026

ABSTRACT

Data Availability Statements (DAS) have become a standard mechanism for promoting transparency and reproducibility in open-access mega journals, yet questions remain about how effectively they support meaningful data sharing in practice and how these practices vary across national research contexts. This study examines data-sharing practices in the medical and health sciences through a comparative analysis of global publications and Korea-affiliated articles in three mega journals indexed in PubMed Central: *PLOS ONE, Scientific Reports* and *BMJ Open* (2020–2024). DAS from 176,145 articles were collected from PubMed Central using an automated pipeline with manual validation and classified into a seven-category typology reflecting levels of data accessibility and reuse. Results indicate that although DAS inclusion increased over time, repository-based data sharing remains limited, while 'data available upon request' continues to be prevalent. Clear differences are observed across journals: *PLOS ONE* shows greater use of repository-based and in-article sharing, whereas *Scientific Reports* and *BMJ Open* rely more heavily on 'data available upon request'. Korea-affiliated articles largely follow global trends, with slightly greater reliance on national public data repositories. Repository use is concentrated among a small number of international multidisciplinary platforms and selected national biomedical databases. The findings reveal a persistent gap between formal DAS compliance and effective data accessibility, indicating the need for clearer, more actionable data-sharing guidance.

1 | Introduction

Data sharing is essential for advancing transparency, reproducibility and collaboration in medical and health research. By enabling independent verification and reuse of data, it strengthens scientific integrity and accountability (Xafis and Labude 2019) and supports evidence-based decision-making in healthcare and public health policy (Hulsen 2020). Shared datasets could inform clinical practice, guide the development of treatment protocols and enhance public health policies, ensuring that such decisions are grounded in accurate and comprehensive evidence (Modi et al. 2025). In this way, data sharing helps bridge research and practice and contributes to improved health outcomes (Hermansen et al. 2022).

Despite growing recognition of its value, data sharing practices remain inconsistent across disciplines, journals and countries. Prior studies have identified persistent barriers, including limited infrastructure, ambiguous journal policies and cultural or institutional reluctance (Federer et al. 2018; Kim et al. 2020). Although open science initiatives and funding policies increasingly emphasise data transparency, it remains unclear how consistently researchers follow through in practice (Melero and Navarro-Molina 2020; Scaffidi et al. 2021).

To enhance transparency and accountability, many journals now require authors to include a Data Availability Statement (DAS), which specifies where and how underlying data can be accessed. Empirical studies have examined the prevalence and

Key Points

- DAS inclusion increased consistently from 2020 to 2024 across all three mega journals, reflecting widespread formal adoption of data availability requirements.
- Repository-based data sharing remains limited, while 'data available upon request' continues to be prevalent, particularly in *Scientific Reports* and *BMJ Open*.
- Korea-affiliated articles largely follow global DAS trends, suggesting that participation in mega journals may standardise data-sharing practices across national contexts, with slightly higher reliance on national public data repositories.
- Repository use is highly concentrated among a small number of international multidisciplinary repositories, alongside selected national biomedical databases, highlighting limited diversification of data-sharing venues.

characteristics of DAS across journals and disciplines, showing the presence of repository-linked DAS is associated with higher citation impact (Colavizza et al. 2020). However, prior research also indicates that the mere presence of a DAS does not necessarily guarantee meaningful data access, raising questions about how these statements are used in practice.

Mega journals have become important venues for examining these patterns because of their open-access policies, broad disciplinary scope and high publication volume (Severin et al. 2020). Their open-access nature ensures immediate accessibility for researchers, educators and policymakers, enabling the rapid dissemination of findings across disciplines (Wakeling et al. 2016). At the same time, concerns have been raised regarding the effectiveness of their data-sharing policies, as review processes often prioritise methodological soundness over novelty, leading to variability in quality and impact (Erfanmanesh and Silva 2019).

While prior studies have largely examined DAS practices at the global level, little attention has been paid to how these practices vary across national research communities. In particular, empirical evidence on data-sharing practices among Korean researchers remains limited, despite increasing policy attention to open science in Korea (Kim et al. 2020; Kim et al. 2023b; Kim et al. 2023a). Examining Korea-affiliated publications in comparison with global patterns provides an opportunity to better understand how national research environments interact with global journal policies.

Therefore, the current study aims to examine data-sharing practices by analysing DAS published between 2020 and 2024 in three mega journals indexed in PubMed Central (PMC)—*PLOS ONE*, *Scientific Reports* and *BMJ Open*. In particular, the study adopts a comparative perspective, examining how patterns of DAS usage, repository selection and access conditions differ between Korea-affiliated publications and the global corpus. The study addresses the following research questions:

- RQ1: What types of data-sharing practices are presented in the DAS of articles published in three mega journals?
- RQ2: How have DAS types changed over time from 2020 to 2024?
- RQ3: How do DAS patterns differ between Korea-affiliated publications and the global dataset across the three mega journals?
- RQ4: Which data repositories and access conditions are most frequently cited, and how do these patterns differ between global authors and Korea-affiliated articles?

By situating Korea-affiliated publications within a broader international context, this study contributes to a better understanding of how journal policies and national research environments shape data-sharing practices in mega journals.

2 | Literature Review

2.1 | Data Sharing in the Medical and Health Sciences

Researchers in the medical and health sciences widely acknowledge the scientific value of data sharing, motivated by goals such as advancing knowledge, fostering collaboration and complying with journal or funder requirements (Hamilton et al. 2023; Hutchings et al. 2020). However, concerns about privacy, confidentiality and data misuse continue to limit open sharing, while incentives such as career recognition remain weak (Dankar 2023; Hamilton et al. 2023).

Despite broad consensus in favour of open science, empirical evidence shows that actual data-sharing practices remain limited. Most data are shared informally, either within institutions or upon personal request, rather than through repositories or journal supplements (Wibowo and Mon 2025). Fewer than 10% of articles make data publicly available, and accessible datasets and shared code remain rare even in journals with mandatory data sharing policies (Hamilton et al. 2023). Barriers include the time required for data curation, lack of training and infrastructure and unclear standards, with clinical research showing more restrictive practices due to ethical constraints (Federer et al. 2015; Hutchings et al. 2020).

In Korea, empirical evidence on data sharing in the medical and health sciences remains limited and has mainly relied on cross-disciplinary surveys rather than large-scale analyses of journal publications. Previous studies suggest growing awareness but limited implementation: half of the researchers engage in partial data sharing, while fully open sharing remains relatively low (around 25%) (Kim, Kim, et al. 2023; Kim, Hwang, et al. 2023a). Structural barriers—such as insufficient infrastructure, weak incentives and concerns over ownership—persist. A national survey by the Science and Technology Policy Institute (STEPI) found that over 60% of Korean researchers never shared data, and only a minority of domestic journals have adopted formal data-sharing policies (Shmagun et al. 2022; Yi et al. 2023).

2.2 | Mega Journals' Data Sharing Policies

Mega journals have played a central role in promoting data transparency through mandatory DAS, although enforcement

varies. *PLOS ONE*, an early adopter, has required all authors to include a DAS specifying the data needed to reproduce results and strongly encourages repository deposition, while allowing in-article or upon-request access when necessary (Federer et al. 2018; Jiao et al. 2024). Nevertheless, only about 16%–20% of articles provide data through public repositories (Federer et al. 2018; Colavizza et al. 2020).

Scientific Reports, part of the Nature Portfolio, follows a similar framework, requiring a DAS and encouraging the repository use and code sharing (Jiao et al. 2024). Authors are expected to make data available in open repositories or domain-specific databases, and to share custom code through a dedicated 'Code Availability' section (Nature Portfolio 2026). Yet, actual repository use remains modest, with many authors still relying on upon-request access.

BMJ Open employs a tiered model (BMJ Publishing Group Ltd. 2026). At Tier 1, authors are required to make data available through repositories upon publication, with controlled access permitted only for legal or ethical reasons. All articles must include a DAS, and registered clinical trials must provide a data-sharing plan at the time of registration. In practice, the overall data-sharing rate remains low—approximately 4.5% across all papers and about 24% among clinical trial articles (Rowhani-Farid and Barnett 2016; Naudet et al. 2018). Overall, while all three mega journals mandate DAS, substantial gaps persist between policy expectations and effective data accessibility.

2.3 | Implementation and Effectiveness of DAS

The DAS specifies how and where underlying data can be accessed, but early studies show limited effectiveness. Research has documented high reliance on 'upon request' statements and low rates of repository-linked data (McDonald et al. 2017; Federer et al. 2018; Jiao et al. 2024). Grant and Hrynaszkiewicz (2018) classified DAS in *Nature* journals into four main types: 'data available upon request', 'included in the manuscript', 'in supplementary materials' or 'publicly available in a repository'. Large-scale analyses later showed that, although DAS adoption rates exceeded 90% in mega journals, most statements still relied on less effective mechanisms such as 'upon request', with only 16%–20% of papers including active repository links (Jiao et al. 2024).

More recent studies have explored the effectiveness of DAS beyond compliance metrics. Colavizza et al. (2020) found that articles with repository links in their DAS received, on average, 25% more citations. Actual data accessibility remains limited: repository links appear in only about 20% of *PLOS ONE* papers and 12% of BMJ papers, while clinical journals report even lower sharing rates. Overall, prior studies consistently point to a persistent policy-practice gap, indicating that the presence of a DAS does not necessarily ensure meaningful data availability.

Although prior studies have examined the prevalence, types and effectiveness of DAS across journals and disciplines, most of this research has focused on global patterns of data-sharing practices. Comparatively little attention has been paid to how these practices vary across national research communities. Examining these differences may provide insights into how national research environments shape data-sharing practices within global publishing platforms such as mega journals. To address this gap, the present study examines DAS in three major mega journals, with particular attention to the comparison between Korea-affiliated publications and the global dataset.

3 | Methods

The current study analysed the DAS from articles published in *PLOS ONE*, *Scientific Reports* and *BMJ Open* between 2020 and 2024. An automated data collection pipeline was developed using the PMC database. DAS texts were extracted via a combination of API queries and web parsing. The following section describes the data sources, extraction procedures and analytical steps.

3.1 | Data Collection

Articles were retrieved from PMC using journal- and date-restricted search queries. Two datasets were constructed: (1) a global corpus of articles published in *PLOS ONE*, *Scientific Reports* and *BMJ Open* between January 1, 2020, and December 31, 2024, and (2) a Korea-affiliated subset consisting of articles from the same journals and time period with at least one author reporting a Korean institutional affiliation. Global articles were retrieved using journal- and date-restricted queries without affiliation constraints. The Korea-affiliated subset was generated using the same queries with an additional affiliation filter ('korea'[Affiliation]). An example of the PMC query used to retrieve the Korea-affiliated subset is shown below:

(*Journal Name* [Journal]) AND 'korea'[Affiliation] AND ('2020/01/01'[PDAT]: '2024/12/31'[PDAT]) AND has_associated_data[Filter].

The *has_associated_data* filter was applied to identify records explicitly containing data availability information. For each record, the PMC ID, bibliographic metadata and DAS text were extracted. Metadata were retrieved via the efetch utility in XML format and parsed to extract article title, authors, affiliations, publication year and Medical Subject Headings (MeSH) terms. Full DAS texts were then collected using a combination of APIs and HTML parsing. Because the analysis relied on PMC full-text/XML records, DAS inclusion in this study refers to DAS text identifiable and extractable from PMC records; DAS not indexed or structurally identifiable in PMC were not included in the analysed dataset.

3.2 | Data Processing and Analysis

All processing and analysis were conducted in Python within a Jupyter Notebook environment. Journal-specific datasets were merged, cleaned and standardised, and records without identifiable DAS were excluded. Some extracted DAS contained

multiple mechanisms (e.g., both a repository link and an 'upon request' clause) or non-standard language that could not be reliably mapped to a single category. To maintain a mutually exclusive classification scheme, each statement was assigned to a single primary category using predefined decision rules during manual validation. Statements that remained non-classifiable after validation were excluded from the category-distribution analyses and reported separately.

DAS texts were then classified using a semi-automated approach. In this study, repository detection adopted a broad operational definition encompassing both author-deposited repositories and externally maintained public data platforms referenced in DAS as mechanisms of data access or reuse. Accordingly, the analysis included not only general-purpose deposition platforms (e.g., Dryad, Figshare, Zenodo, OSF) but also curated public databases and national data providers frequently cited in DAS (e.g., NCBI, SEER, World Bank, NHIS Korea).

Repository and data platform names were identified using keyword-based detection combined with an expanded dictionary of frequently used repositories, databases and data platforms. Because DAS frequently contained heterogeneous naming conventions, abbreviations, URLs and institutional references, all automatically identified entities were subsequently manually validated and normalised using a custom Tkinter-based interface. This process enabled the correction of ambiguous or misclassified cases, the merging of synonymous platform names and the removal of false positives. This human-in-the-loop validation process improved the reliability and consistency of the final coded dataset.

4 | Results

A total of 478,844 full-text articles published between 2020 and 2024 were retrieved from PMC, including *PLOS ONE* ($n = 175{,}702$), *Scientific Reports* ($n = 247{,}098$) and *BMJ Open* ($n = 56{,}044$). Among these, 176,145 articles (36.79%) contained DAS text identifiable from PMC full-text/XML records. DAS inclusion varied substantially by journal. *Scientific Reports* (44.67%) exhibited the highest inclusion rate, followed by *PLOS ONE* (35.45%) and *BMJ Open* (17.94%). Across all three journals, DAS inclusion steadily increased from 2020 to 2024, from 19.95% ($n = 25{,}384$) to 52.35% ($n = 38{,}629$).

In this dataset, 26,755 articles (5.59% of the total) included at least one author affiliated with a Korean institution: 10,139 in *PLOS ONE*, 16,049 in *Scientific Reports* and 567 in *BMJ Open*. Among these, 9414 articles (35.19%) included a DAS. Similar to the global trend, DAS inclusion among Korea-affiliated articles increases over time—from 22.18% ($n = 1330$) in 2020 to 71.19% ($n = 2401$) in 2024. Analyses focus on articles with an extracted DAS mapped to one of seven categories. Non-classifiable statements were excluded from category analyses.

4.1 | Definition and Classification of DAS Types

Building upon previous large-scale DAS typologies (Federer et al. 2018; Jiao et al. 2024), this study employed a refined seven-category classification framework to capture the variations in data accessibility and reuse. Each DAS was categorised by data availability, location and accessibility level (Table 1).

Table 2 summarises the overall distribution of DAS types across the three mega journals, comparing the global dataset ('All') with articles including at least one Korean-affiliated author ('Korea'). The global dataset includes Korean-affiliated articles, which constitute only about 6% of the total corpus. DAS types are ordered by descending frequency in the global dataset. Across all journals, upon request, in article/supplement and repository are the most prevalent, together accounting for more than 90% of all statements.

However, the relative prevalence of DAS varies substantially by journals. *PLOS ONE* relies on most heavily on in article/supplement (47.70%) and repository (33.95%), with limited use of upon request (9.90%). *Scientific Reports* is dominated by upon request (60.44%), followed by repository (23.19%) and in article/supplement (9.39%). *BMJ Open* shows a similar reliance on upon request (58.61%), while repository (10.29%) and in article/supplement (7.60%) remain comparatively low.

The Korean-affiliated subset mirrors these global patterns but shows notable journal-specific differences. In *PLOS ONE*, Korea-affiliated articles show higher use of in article/supplement statements (48.76%) and lower reliance on upon request (10.80%). *Scientific Reports*, by contrast, remains strongly dominated by upon request (68.62%) with limited use of in article/supplement (7.15%). Repository use remains moderate across journals—highest in *PLOS ONE* (32.52%), followed by *BMJ Open* (23.08%) and *Scientific Reports* (17.80%).

BMJ Open represents the most apparent divergence from the global trend in the Korea-affiliated subset, with a substantially higher reliance on repository-based sharing (Global: 10.29% vs. Korea: 23.08%). This contrast is consistent with the possibility that *BMJ Open*'s more structured, clinically oriented data-sharing framework exerts a stronger influence on data-sharing decisions in Korea.

Less common DAS types—unknown, no dataset, not shared/restricted and public data use—each account for less than 7% of statements. Because the 'unknown' category lacks an identifiable sharing mechanism, it was excluded from all subsequent analyses.

Figure 1 visualises the distribution patterns of DAS types across the three mega journals, enabling direct comparison between the global dataset and the Korea-affiliated subset. The radar charts show distinct journal-level profiles, indicating that the relative prevalence of DAS types differs substantially across journals. Within each journal, the overall patterns observed in Korea-affiliated publications generally resemble those of the global dataset (Figure 1a), although several differences are visible. In *Scientific Reports*, Korea-affiliated publications show a stronger reliance on 'upon request' statements compared with the global dataset (Figure 1b). By contrast, in

TABLE 1 | DAS type definitions and examples.

DAS types	Definitions	Example statements
Upon request	Data are not publicly shared but can be provided by the corresponding author upon reasonable request.	'The data sets used for this manuscript are available from the corresponding author upon reasonable request'.
Repository	Data are deposited in an external, public or domain-specific repository with a persistent identifier such as a DOI, URL or accession number.	'All relevant data are available at https://github.com/REPO_NAME'.[a]
Public data use	External publicly available datasets were reused and properly cited.	'Data are available in a public, open-access repository. The Korea Health Panel Survey data used in this article are available in LINK'.[a]
In article/supplement	Data are provided directly within the manuscript text, figures, tables or supplementary files.	'All data relevant to the study are included in the article or uploaded as online supplemental information'.
Not shared/restricted	Data access is restricted and cannot be provided due to legal, contractual, institutional, ethical or privacy-related constraints.	'No data are available. Due to the sensitive nature of the data and the potential for participants to be identified, the dataset is not publicly available'.
No dataset	No dataset was generated or analysed in the study.	'No data have been generated or analysed in this manuscript'.
Unknown	An ambiguous explanation was illustrated in the Data Availability Statement section.	'Data are available in a public, open-access repository'. 'Data will be available upon the study's completion'.

[a]URLs and project identifiers have been anonymised or redacted.

TABLE 2 | Distribution of DAS types by journal.

		PLOS ONE		*Scientific Reports*		*BMJ Open*		Total	
DAS types		***N***	**%**	***N***	**%**	***N***	**%**	***N***	**%**
Upon request	All	7774	9.90%	52,946	60.44%	5891	58.61%	66,611	37.82%
	Korea	415	10.80%	3735	68.62%	64	49.23%	4214	44.76%
Repository	All	26,644	33.95%	20,313	23.19%	1034	10.29%	47,991	27.25%
	Korea	1249	32.52%	969	17.80%	30	23.08%	2248	23.88%
In article/supplement	All	37,438	47.70%	8224	9.39%	764	7.60%	46,426	26.36%
	Korea	1873	48.76%	389	7.15%	7	5.38%	2269	24.10%
Unknown	All	4550	5.80%	5361	6.12%	507	5.04%	10,418	5.91%
	Korea	234	6.09%	287	5.27%	6	4.62%	527	5.60%
No dataset	All	1045	1.33%	304	0.35%	1488	14.80%	2837	1.61%
	Korea	26	0.68%	25	0.46%	14	10.77%	65	0.69%
Not shared/restricted	All	827	1.05%	316	0.36%	358	3.56%	1501	0.85%
	Korea	39	1.02%	32	0.59%	9	6.92%	80	0.85%
Public data use	All	208	0.27%	143	0.16%	10	0.10%	361	0.20%
	Korea	5	0.13%	6	0.11%	0	0.00%	11	0.12%

BMJ Open, repository-based sharing appears relatively more prominent in Korea-affiliated publications (Figure 1c). These visual patterns reinforce the journal-level differences reported in Table 2 while highlighting variation in the relative balance of DAS types across journals.

4.2 | Five-Year Trends in DAS Types (2020–2024): Global vs. Korea

Figure 2 presents longitudinal trends in DAS types across journals from 2020 to 2024. Both the global and the Korean datasets

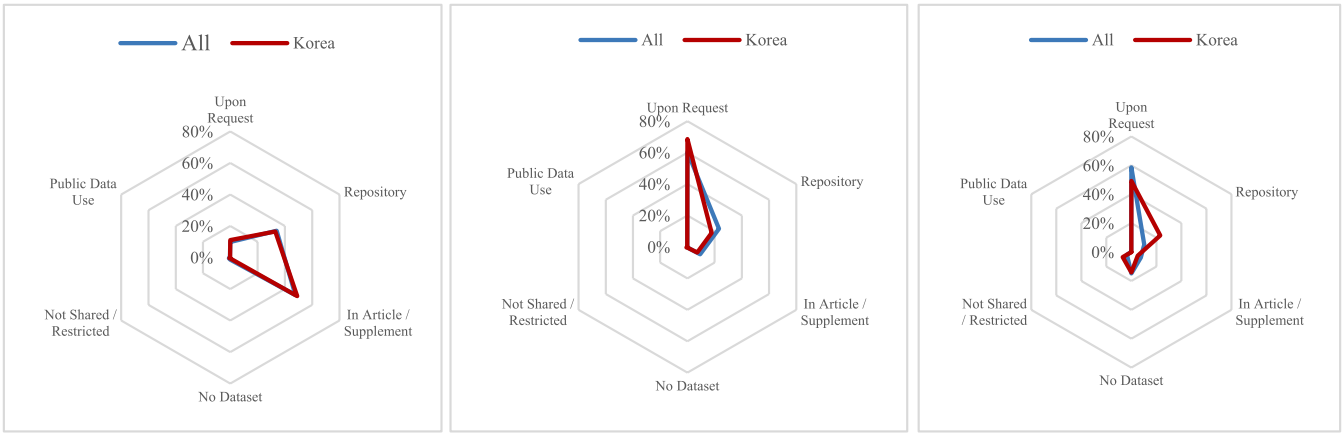

FIGURE 1 | Distribution patterns of DAS types across mega journals. (a) *PLOS ONE*, (b) *Scientific Reports* and (c) *BMJ Open*.

reveal a gradual decline in the upon request type and a corresponding increase in repository-based sharing.

In *PLOS ONE* (Figure 2a), data-sharing practices gradually shift over time. In article/supplement declines from 54.27% to 44.80%, while Repository use increases from 31.11% to 36.35%. Upon request remains consistently low (approximately 9%–11%). The Korean-affiliated subset (Figure 2b) follows a similar pattern, with in article/supplement decreasing from 60.23% to 45.37% and repository use increasing from 25.95% to 35.41%.

In *Scientific Reports*, both the global dataset (Figure 2c) and the Korea-affiliated subset (Figure 2d) display highly stable patterns. Upon request dominates throughout the period (55.39%–63.10% globally, 62.02%–71.98% for Korea-affiliated articles), while repository and in article/supplement remain comparatively low and show minimal year-to-year variation. A modest increase in in article/supplement is observed in the Korea subset, but its overall share remains small.

For *BMJ Open*, the global dataset (Figure 2e) also shows strong stability, with upon request consistently accounting for approximately 56%–61% of statements, while other categories remain low. The Korea-affiliated subset (Figure 2f) exhibits greater fluctuation due to small sample sizes, including an inflated in article/supplement share in 2020, reflecting that only a single Korea-affiliated article with a DAS was published that year, resulting in a proportion of 100%. From 2021 onward, the distribution stabilises, and by 2024, repository and upon request emerge as the dominant categories, indicating relatively greater reliance on repository-based sharing than in the global dataset.

4.3 | Repository Usage Patterns

Repository usage follows a pronounced long-tail distribution. Across all DAS, 3843 unique repositories/data platforms were identified globally, and 327 in the Korea-affiliated subset. Of these, 332 (global) and 26 (Korea) appeared at least five times and were included in the analysis of frequently used platforms, while the majority appeared only sporadically.

Figure 3 presents the top 20 repositories. Globally (Figure 3a), repository use is highly concentrated: GitHub, NCBI, OSF, Figshare and Zenodo together account for 59.69% of all repository mentions. The Korea-affiliated subset (Figure 3b) shows a similar but more localised pattern, with NCBI, GitHub, NHIS Korea, KNHANES, Figshare and OSF accounting for just over half of all mentions. The prominence of national health databases (e.g., NHIS Korea, KNHANES) highlights a dual reliance on international platforms and domestic public data infrastructures.

Across journals, *Scientific Reports* exhibits the highest level of repository-based sharing, whereas *BMJ Open* shows the lowest. Although PLOS ONE includes the largest absolute number of articles with repository-linked DAS, its repository use is distributed across a broader range of platforms, resulting in a lower concentration among the most frequently cited repositories. In contrast, *Scientific Reports* shows a higher concentration of repository mentions among a smaller number of dominant platforms.

4.4 | Repository Characteristics—Governance, Scope and Access

To examine the structural pattern in repositories referenced in DAS, repositories appearing in Figure 3 were classified by governance (international, national or institutional), disciplinary scope (disciplinary or multidisciplinary), and access model (upload-available or download-enabled).

Table 3 presents the repositories appearing in the top 20 lists for the global ('All') and Korea-affiliated ('Korea') datasets. In total, 28 distinct repositories were identified. By governance, repositories were predominantly international ($n = 22{,}123$, 75.38%), followed by national ($n = 7224$, 24.62%), with institutional repositories accounting for only a small fraction ($n = 912$, 3.11%). In terms of scope, the multidisciplinary repositories ($n = 27{,}349$, 93.19%) overwhelmingly dominated, while disciplinary repositories accounted for a small share ($n = 1998$, 6.81%). Regarding access characteristics, most repositories support user uploads ($n = 27{,}402$, 93.37%) and data downloads ($n = 29{,}347$, 100.00%). Overall, the repository landscape is strongly shaped by

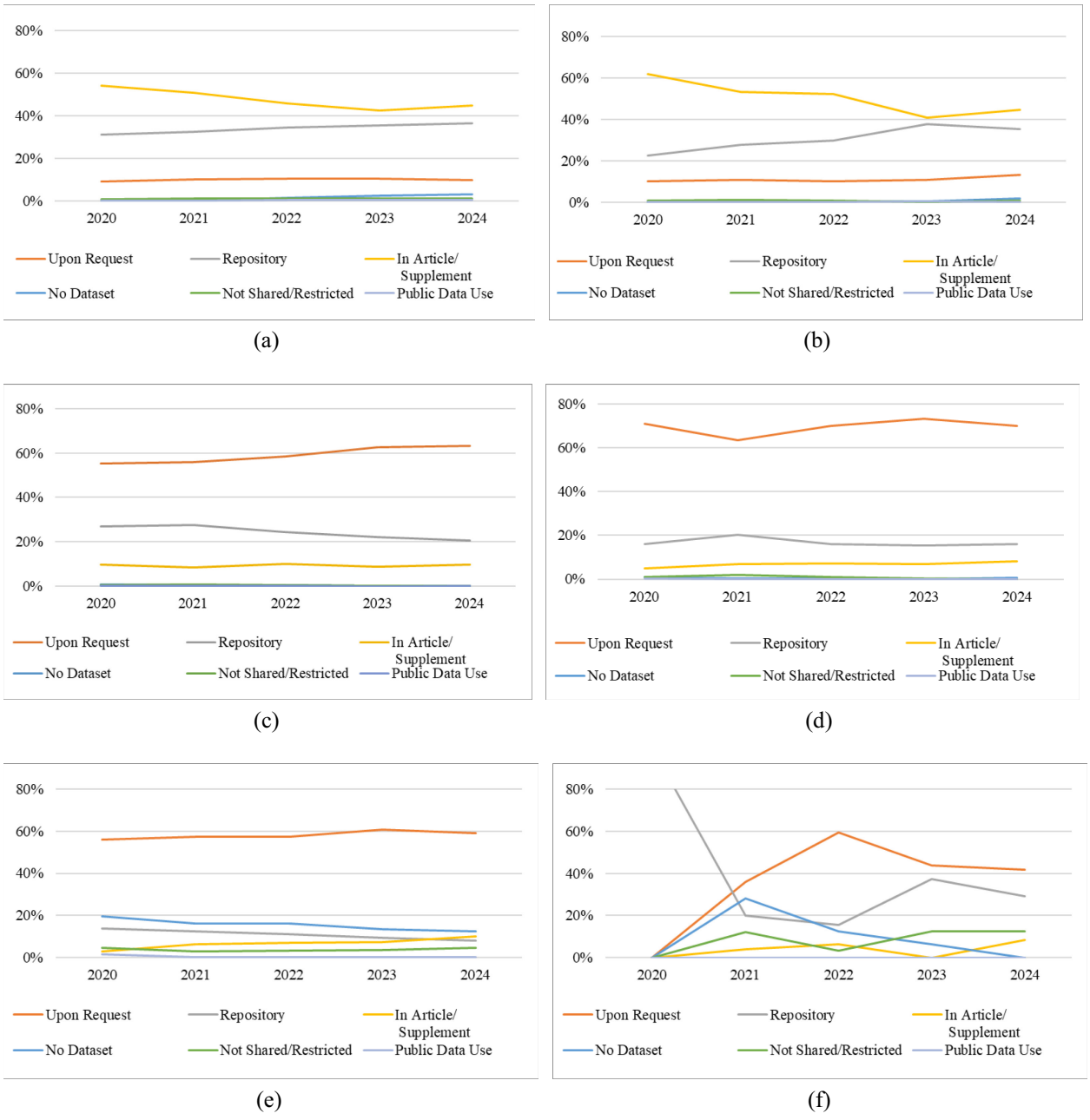

FIGURE 2 | Distribution of DAS types in global and Korean subsets, 2020–2024. (a) DAS trends in *PLOS ONE*: all, (b) DAS trends in *PLOS ONE*: Korea, (c) DAS trends in *Scientific Reports*: all, (d) DAS trends in *Scientific Reports*: Korea, (e) DAS Trends in *BMJ Open*: all, (f) DAS trends in *BMJ Open*: Korea.

international, multidisciplinary and upload-enabled platforms, reflecting the infrastructure of contemporary open science in the medical and health sciences.

The most frequently cited repositories—such as GitHub, OSF, Figshare, Zenodo, Dryad, Dataverse and Mendeley Data—are international, multidisciplinary platforms that facilitate user-driven data and code deposition. Alongside these, a smaller but important group of national repositories (e.g., NCBI, CDC, KDCA, NHS and NHIS Korea) appears frequently, particularly in biomedical and public health research. These repositories often function as national data custodians, emphasising curated or regulated data access rather than open public contribution.

Disciplinary repositories are largely concentrated in genomics and biomedical domains (e.g., GenBank, GEO, SRA, ENA, DDBJ, PDB, MG-RAST), reflecting long-standing norms that encourage or mandate deposition of structured biological data. In contrast, institutional repositories (e.g., university-based platforms) appear infrequently, suggesting that journal- and domain-level repositories play a more central role than institutional archiving systems in medical and health science publishing.

4.5 | Repository Characteristics by Journals

Tables 4–6 summarise repository characteristics by governance, disciplinary scope and access mechanism across the three

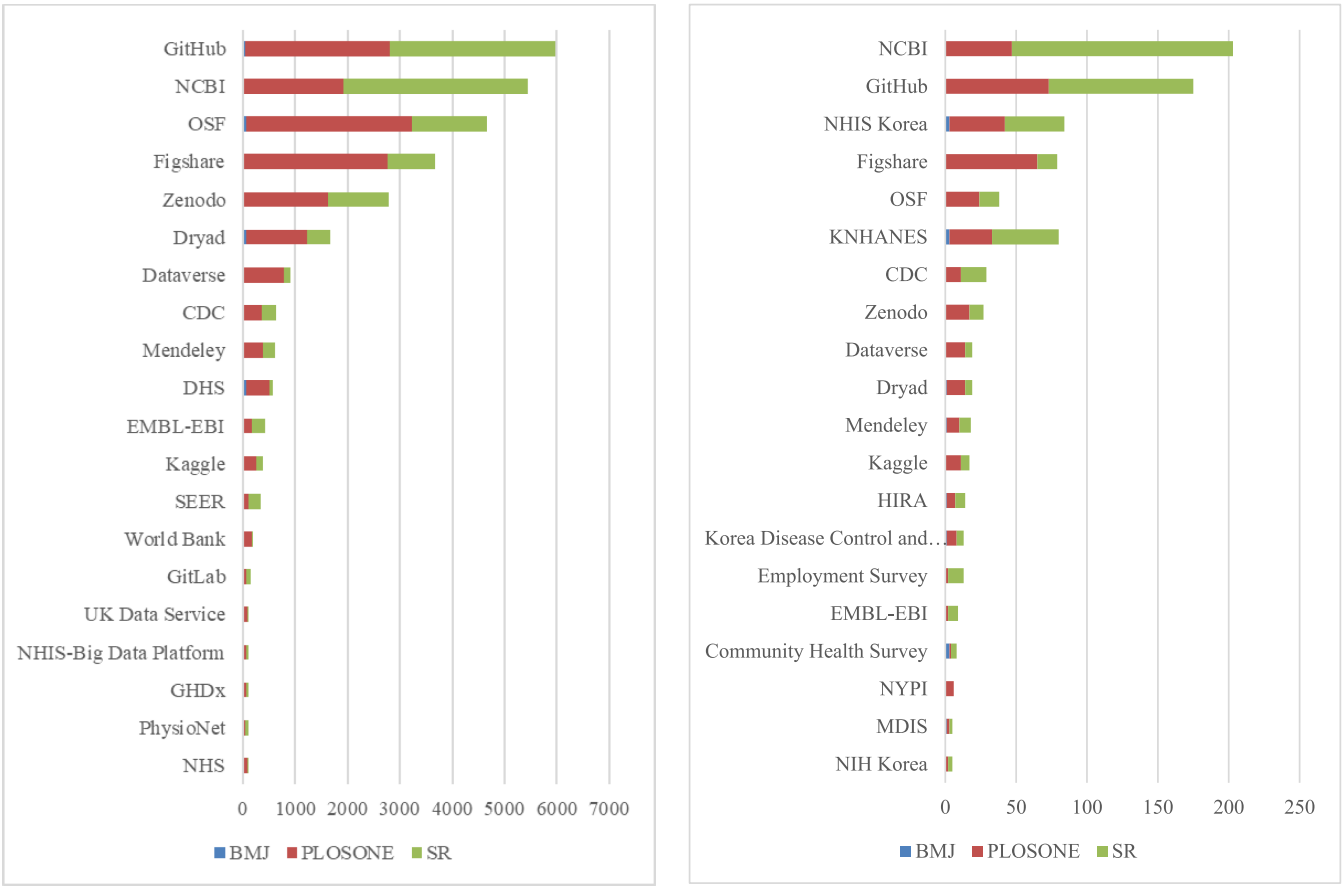

FIGURE 3 | Top repositories and data platforms referenced in DAS. (a) Repositories: all and (b) repositories: Korea.

journals, for both the global ('All') and Korea-affiliated ('Korea') datasets.

Across journals, international repositories dominate, with *PLOS ONE* and *Scientific Reports* showing particularly strong reliance on international platforms, while institutional repositories are negligible (< 5%) (Table 4). National repositories constitute a substantial secondary category, most notably in *Scientific Reports* and in the Korea-affiliated subset of *BMJ Open*, where they account for the majority of repository mentions.

Clear journal-level differences also emerge in disciplinary scope (Table 5). In the global dataset, multidisciplinary repositories predominate in *PLOS ONE* and *Scientific Reports*, whereas *BMJ Open* relies more heavily on disciplinary repositories. This pattern is amplified in the Korea-affiliated subset, where disciplinary repositories account for a large majority of *BMJ Open* citations, while multidisciplinary platforms remain dominant in the other two journals.

Access mechanism was measured as the dominant mode implied by the DAS: deposit-oriented (upload) or custodian-oriented (download/access). Table 6 shows mutually exclusive distributions. With respect to access mechanisms (Table 6), *PLOS ONE* and *Scientific Reports* show a relatively balanced distribution between upload-oriented and download-oriented repositories, whereas *BMJ Open* exhibits a stronger reliance on download-oriented repositories in both the global and Korea-affiliated datasets. Together, these patterns indicate that repository access patterns vary systematically by journal, reflecting differences in editorial scope, disciplinary focus and data-sharing expectations.

5 | Discussion

5.1 | Differences in Data-Sharing Practices Across Mega Journals

This study demonstrates persistent differences in data-sharing practices across the three mega journals despite their shared commitment to open access and the requirement of mandatory DAS. Although prior studies show that DAS sections are present in more than 90% of articles following the introduction of mandatory journal policies (Jiao et al. 2024), these figures primarily reflect policy compliance rather than the extent to which the statements provide actionable access to data. In contrast, our results DAS text identifiable and extractable from PMC full-text/XML records rather than overall compliance with journal DAS policies. Under this operational definition, repository-based sharing remains limited, while conditional access (e.g., 'upon request') continues to be common (Federer et al. 2018; Jiao et al. 2024; Colavizza et al. 2020). The proportion of DAS varied substantially by journal, with *Scientific Reports* (44.67%) showing the highest rate, followed by *PLOS ONE* (35.45%) and *BMJ Open* (17.94%).

TABLE 3 | Classification of repositories cited in DAS by governance, disciplinary scope and access mechanisms.

Repository	*N*	International	National	Institutional	Disciplinary	Multi-disciplinary	Upload-enabled	Download-enabled
GitHub	5980	✓	—	—	—	✓	✓	✓
National Center for Biotechnology Information (NCBI)	5436	—	✓	—	—	✓	✓	✓
Open Science Framework (OSF)	4656	✓	—	—	—	✓	✓	✓
Figshare	3667	✓	—	—	—	✓	✓	✓
Zenodo	2781	✓	—	—	—	✓	✓	✓
Dryad	1667	✓	—	—	—	✓	✓	✓
Dataverse	912	✓	—	✓	—	✓	✓	✓
Centers for Disease Control and Prevention (CDC)	664	—	✓	—	—	✓	—	✓
Mendeley Data	620	✓	—	—	—	✓	✓	✓
Demographic and Health Surveys (DHS)	584	✓	—	—	✓	—	—	✓
European Molecular Biology Laboratory—European Bioinformatics Institute (EMBL—EBI)	435	✓	—	—	✓	—	✓	✓
Kaggle	377	✓	—	—	—	✓	✓	✓
Surveillance, Epidemiology and End Results Program (SEER)	339	—	✓	—	✓	—	✓	✓
World Bank	185	✓	—	—	—	✓	—	✓
GitLab	146	✓	—	—	—	✓	✓	✓
National Health Insurance Service Korea (NHIS Korea)	139	—	✓	—	—	✓	—	✓
PhysioNet	128	—		—	✓	——	✓	✓

(Continues)

TABLE 3 | (Continued)

Repository	*N*	International	National	Institutional	Disciplinary	Multi-disciplinary	Upload-enabled	Download-enabled
UK Data Service	119	—	✓	—	✓	—	✓	✓
Global Health Data Exchange (GHDx)	113	✓		—	—	✓	—	✓
National Institutes of Health Korea (NIH Korea)	112	—	✓	—	—	✓	✓	✓
Korea National Health and Nutrition Examination Survey (KNHANES)	95	—	✓	—	✓	—	—	✓
National Health Service (NHS)	93	—	✓	—	✓	—	—	✓
Korea Disease Control and Prevention Agency (KDCA)	59	—	✓	—	✓	—	—	✓
Health Insurance Review & Assessment (HIRA)	22	—	✓	—	✓	—	—	✓
Employment Survey	19	—	✓	—	✓	—	——	✓
Community Health Survey	18	—	✓	—	✓	—	—	✓
National Youth Policy Institure (NYPI)	7	—	✓	——	✓	—	—	✓
The Management Development Institute of Singapore (MDIS)	6	—	✓	—	—	✓	—	✓

Longitudinal analysis further reveals a clear increase in identifiable DAS records over the past five years, indicating growing normalisation of data transparency requirements in medical and health research. However, journal-specific patterns remained relatively stable: *PLOS ONE* relied more on repository-based and in-article sharing, whereas *Scientific Reports* and *BMJ Open* were dominated by 'upon request' statements. These differences in DAS patterns across journals may reflect variations in editorial emphasis and journal policies, as well as potential disciplinary differences in the research communities publishing in each journal. For example, journals that publish a larger proportion of studies involving human subjects, clinical data or regulated datasets may show greater reliance on conditional access statements due to ethical and legal constraints on public data sharing (Taichman et al. 2017). Future research could further examine how differences in DAS patterns relate to variation in

TABLE 4 | Repository category distribution by governance.

	PLOS ONE		*Scientific Reports*		*BMJ Open*	
Governance	***N***	**%**	***N***	**%**	***N***	**%**
Repositories: All						
International	13,697	79.47%	8072	64.37%	354	73.14%
National	2771	16.08%	4340	34.61%	113	23.35%
Institutional	767	4.45%	128	1.02%	17	3.51%
Total	17,235	100%	12,540	100%	484	100%
Repositories: Korea						
International	544	64.00%	255	40.48%	7	33.33%
National	283	33.29%	369	58.57%	14	66.67%
Institutional	23	2.71%	6	0.95%	0	0.00%
Total	850	100%	630	100%	21	100%

TABLE 5 | Repository category distribution by disciplinary scope.

	PLOS ONE		*Scientific Reports*		*BMJ Open*	
Scope	***N***	**%**	***N***	**%**	***N***	**%**
Repositories: All						
Disciplinary	1050	6.38%	810	6.53%	138	29.55%
Multidisciplinary	15,418	93.62%	11,602	93.47%	329	70.45%
Total	16,468	100%	12,412	100%	467	100%
Repositories: Korea						
Disciplinary	170	20.56%	150	24.04%	14	66.67%
Multidisciplinary	657	79.44%	474	75.96%	7	33.33%
Total	827	100%	624	100%	21	100%

TABLE 6 | Repository category distribution by access mechanism.

	PLOS ONE		*Scientific Reports*		*BMJ Open*	
Governance	***N***	**%**	***N***	**%**	***N***	**%**
Repositories: All						
Upload-enabled	15,276	48.12%	11,842	48.82%	284	37.82%
Download-enabled	16,468	51.88%	12,412	51.18%	467	62.18%
Total	31,744	100%	24,254	100%	751	100%
Repositories: Korea						
Upload-enabled	634	43.39%	460	42.44%	4	16.00%
Download-enabled	827	56.61%	624	57.56%	21	84.00%
Total	1461	100%	1084	100%	25	100%

data sensitivity, governance constraints and data types in medical and health research.

5.2 | Global and Korea-Affiliated Patterns: Similarities and Subtle Divergences

Contrary to expectations of substantial national variation, Korea-affiliated articles largely mirror global patterns across journals and over time. The same dominant DAS types and repositories/data platforms appear in both datasets, indicating that participation in mega journals may standardise data-sharing practices across national contexts. Because mega journals operate under common editorial policies and transparency requirements, authors from different research environments may adopt similar approaches to describing data availability when publishing in these venues.

Nevertheless, some differences are observable. Korea-affiliated publications show slightly higher reliance on certain repositories and modest variation in DAS types in specific journals. These patterns may reflect characteristics of the Korean research environment, including evolving open-science policies (Shmagun et al. 2022) and the availability of national data infrastructures (Kim, Kim, et al. 2023; Kim, Hwang, et al. 2023a). One notable difference is that Korea-affiliated articles in *BMJ Open* show a relatively higher proportion of repository-based sharing compared to the global dataset, with greater use of disciplinary and national repositories, suggesting that local data infrastructures may complement international platforms in managing sensitive or regulated datasets. Future research could further investigate how national policy environments, funding mandates and research data ecosystems influence repository selection in international publishing contexts.

5.3 | Repository Use and Journal-Specific Infrastructure

Journal-level differences also extend to repository characteristics. Repository use follows a pronounced long-tail distribution, with a small number of international, multidisciplinary platforms—such as GitHub, NCBI, OSF, Figshare and Zenodo—accounting for most citations. These platforms provide flexible infrastructure capable of accommodating diverse data types across disciplines, which may explain their dominant presence across the journals analysed.

At the same time, repository patterns differ by journal. Interpreted together with Table 2 and Figure 3, the journals differ both in the prevalence of repository-based statements and in the concentration of repository mentions across platforms, with *PLOS ONE* showing the highest share of repository-based statements and *Scientific Reports* exhibiting stronger concentration on a smaller set of dominant repositories. This contrast suggests that repository use reflects not only author preference but also disciplinary conventions and data-sharing infrastructures associated with particular research communities. *BMJ Open*'s relatively stronger reliance on disciplinary and download-oriented repositories highlights the influence of domain-specific infrastructures, particularly in clinical and biomedical research where data governance and ethical considerations often shape data accessibility.

6 | Limitations

This study has several limitations. First, articles were classified as Korea-affiliated if at least one author listed a Korean institutional affiliation. Consequently, the Korea-affiliated subset (approximately 6% of the dataset) includes many internationally co-authored papers and does not represent exclusively national research practices. While this approach allows comparison with the global corpus, future studies could adopt a more fine-grained, country-level design to systematically examine national differences across multiple countries. Second, DAS often describe both original data deposition and the reuse of external public datasets, and these practices are difficult to distinguish reliably through large-scale text analysis. Because this study relied on author-provided DAS text, the analysis adopted a broad operational definition that included both author-deposited repositories and publicly accessible data platforms referenced in DAS. While this approach enabled consistent large-scale analysis, it may blur the distinction between data sharing and data reuse in some cases.

7 | Conclusion

This study provides a large-scale, longitudinal analysis of the DAS across three leading open-access mega journals, offering empirical insight into how data-sharing policies are implemented in practice in the medical and health sciences. Despite widespread adoption of mandatory DAS and steady growth in their inclusion over time, effective data sharing—particularly through public repositories—remains uneven and strongly shaped by journal-specific norms rather than national research contexts. By focusing on the substantive content of DAS, this study demonstrates that formal compliance with DAS requirements does not necessarily translate into effective data accessibility. The persistence of journal-specific patterns and the continued dominance of conditional access statements indicate that data-sharing practices remain shaped by publishing norms and infrastructural arrangements. These findings provide empirical evidence of a structural gap between policy expectations and actual data-sharing practices in medical and health research, offering a foundation for future research examining how data-sharing policies are interpreted and enacted across journals and disciplines.

Author Contributions

Sanghee Oh: conceptualization, Methodology, Formal analysis, Data curation, Writing – original draft, Writing – review and editing. **Yunseo Park:** investigation, Data curation, Formal analysis, Writing – review and editing. **Seyun Sim:** investigation, Data curation, Formal analysis, Writing – review and editing. All authors have read and agreed to the published version of the manuscript.

Funding

This study was supported by the SKKU Academic Research Support Program, Sungkyunkwan University, 2024.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data supporting this study are publicly available in Zenodo at https://zenodo.org/records/20786170. The deposited files include the full analysis dataset and a Korea-affiliated subset, with article-level information on PMCID, journal, publication year, extracted DAS text, DAS classification, and detected repository name where applicable.

References

BMJ Publishing Group Ltd. 2026. *Data Sharing—Author Hub Policies*. BMJ Publishing Group Ltd. Retrieved November 12, 2025. https://authors.bmj.com/policies/data-sharing/.

Colavizza, G., I. Hrynaszkiewicz, I. Staden, K. Whitaker, and B. McGillivray. 2020. "The Citation Advantage of Linking Publications to Research Data." *PLoS One* 15, no. 4: e0230416.

Dankar, F. K. 2023. "Practices and Challenges in Clinical Data Sharing." *arXiv*. arXiv:2304.06509. https://doi.org/10.48550/arXiv.2304.06509.

Erfanmanesh, M., and J. Silva. 2019. "Is the Soundness-Only Quality Control Policy of Open Access Mega Journals Linked to a Higher Rate of Published Errors?" *Scientometrics* 120: 917–923. https://doi.org/10.1007/s11192-019-03153-5.

Federer, L. M., C. W. Belter, D. J. Joubert, et al. 2018. "Data Sharing in PLOS ONE: An Analysis of Data Availability Statements." *PLoS One* 13, no. 5: e0194768.

Federer, L. M., Y.-L. Lu, D. J. Joubert, J. Welsh, and B. Brandys. 2015. "Biomedical Data Sharing and Reuse: Attitudes and Practices of Clinical and Scientific Research Staff." *PLoS One* 10, no. 6: e0129506. https://doi.org/10.1371/journal.pone.0129506.

Grant, R., and I. Hrynaszkiewicz. 2018. "The Impact on Authors and Editors of Introducing Data Availability Statements at Nature Journals." *International Journal of Digital Curation* 13, no. 1: 195–203. https://doi.org/10.2218/ijdc.v13i1.614.

Hamilton, D. G., K. Hong, H. Fraser, A. Rowhani-Farid, F. Fidler, and M. J. Page. 2023. "Prevalence and Predictors of Data and Code Sharing in the Medical and Health Sciences: Systematic Review With Meta-Analysis of Individual Participant Data." *BMJ* 382: e075767.

Hermansen, A., D. A. Regier, and S. Pollard. 2022. "Developing Data Sharing Models for Health Research With Real-World Data: A Scoping Review of Patient and Public Preferences." *Journal of Medical Systems* 46, no. 12: 86.

Hulsen, T. 2020. "Sharing Is Caring—Data Sharing Initiatives in Healthcare." *International Journal of Environmental Research and Public Health* 17: 3046. https://doi.org/10.3390/ijerph17093046.

Hutchings, E., M. Loomes, P. Butow, and F. Boyle. 2020. "A Systematic Literature Review of Researchers' and Healthcare Professionals' Attitudes Towards the Secondary Use and Sharing of Health Administrative and Clinical Trial Data." *Systematic Reviews* 9: 240. https://doi.org/10.1186/s13643-020-01485-5.

Jiao, C., K. Li, and Z. Fang. 2024. "Data Sharing Practices Across Knowledge Domains: A Dynamic Examination of Data Availability Statements in PLOS ONE Publications." *Journal of Information Science* 50, no. 3: 673–689.

Kim, J., H. Hwang, Y. Jung, S. Cho, and T. Seo. 2023a. "Data Sharing Attitudes and Practices of Researchers in Korean Government Research Institutes: A Survey-Based Descriptive Study." *Science Editing* 10: 71–77. https://doi.org/10.6087/kcse.299.

Kim, J., C. Kim, K. Kim, et al. 2023b. "Scalable Infrastructure Supporting Reproducible Nationwide Healthcare Data Analysis Toward FAIR Stewardship." *Scientific Data* 10: 1–9. https://doi.org/10.1038/s41597-023-02580-7.

Kim, J., S. Kim, H.-M. Cho, J. H. Chang, and S. Y. Kim. 2020. "Data Sharing Policies of Journals in Life, Health, and Physical Sciences Indexed in Journal Citation Reports SCIE and Scopus." *PeerJ* 8: e9924. https://doi.org/10.7717/peerj.9924.

McDonald, L., A. Schultze, A. Simpson, S. Graham, R. Wasiak, and S. V. Ramagopalan. 2017. "A Review of Data Sharing Statements in Observational Studies Published in the BMJ: A Cross-Sectional Study." *F1000Research* 6: 1708.

Melero, R., and C. Navarro-Molina. 2020. "Researchers' Attitudes and Perceptions Towards Data Sharing and Data Reuse in the Field of Food Science and Technology." *Learned Publishing* 33: 163–179. https://doi.org/10.1002/leap.1287.

Modi, N. D., L. X. Li, J. M. Logan, et al. 2025. "The State of Individual Participant Data Sharing for the Highest-Revenue Medicines." *Clinical Trials* 22, no. 2: 170–177.

Nature Portfolio. 2026. *Reporting Standards and Availability of Data, Materials, Code and Protocols*. Springer Nature. Retrieved November 12, 2025. https://www.nature.com/nature-portfolio/editorial-policies/reporting-standards#availability-of-materials.

Naudet, F., C. Sakarovitch, P. Janiaud, et al. 2018. "Data Sharing and Reanalysis of Randomized Controlled Trials in Leading Biomedical Journals With a Full Data Sharing Policy: Survey of Studies Published in the BMJ and PLOS Medicine." *BMJ* 360: k400. https://doi.org/10.1136/bmj.k400.

Rowhani-Farid, A., and A. Barnett. 2016. "Has Open Data Arrived at the British Medical Journal (BMJ)? An Observational Study." *BMJ Open* 6: e011784. https://doi.org/10.1136/bmjopen-2016-011784.

Scaffidi, M., K. Elsolh, J. Li, et al. 2021. "Do Authors of Research Funded by the Canadian Institutes of Health Research Comply With Its Open Access Mandate?: A Meta-Epidemiologic Study." *PLoS One* 16: e0256577. https://doi.org/10.1371/journal.pone.0256577.

Severin, A., M. Egger, M. P. Eve, and D. Hürlimann. 2020. "Discipline-Specific Open Access Publishing Practices and Barriers to Change: An Evidence-Based Review." *F1000Research* 7: 1925.

Shmagun, H., J. Shim, K. Choi, S. Shin, J. Kim, and C. Oppenheim. 2022. "Korea's National Approach to Open Science: Present and Possible Future." *Journal of Information Science* 50: 766–785. https://doi.org/10.1177/01655515221107336.

Taichman, D., P. Sahni, A. Pinborg, et al. 2017. "Data Sharing Statements for Clinical Trials: A Requirement of the International Committee of Medical Journal Editors." *Annals of Internal Medicine* 167: 63–65. https://doi.org/10.7326/m17-1028.

Wakeling, S., P. Willett, C. Creaser, J. Fry, S. Pinfield, and V. Spezi. 2016. "Open-Access Mega-Journals: A Bibliometric Profile." *PLoS One* 11: e0165359. https://doi.org/10.1371/journal.pone.0165359.

Wibowo, M., and L. Mon. 2025. "Investigating the Practices and Preferences of Health Scholars in Sharing Open Research Data." *PLoS One* 20: e0313644. https://doi.org/10.1371/journal.pone.0313644.

Xafis, V., and M. K. Labude. 2019. "Openness in Big Data and Data Repositories: The Application of an Ethics Framework for Big Data in Health and Research." *Asian Bioethics Review* 11, no. 3: 255–273.

Yi, H., Y. Jung, H. Hwang, and S. Cho. 2023. "Korean Scholarly Journal Editors' and Publishers' Attitudes Towards Journal Data Sharing Policies and Data Papers (2023): A Survey-Based Descriptive Study." *Science Editing* 10: 141–148. https://doi.org/10.6087/kcse.316.