A Decade of Progress: Insights of Open Data Practices in Biosciences at the University of Edinburgh

Authors:

Haya Deeb[1], Hwee Yun Wong[1], Trisha Usman[1], Megan A. M. Kutzer[2], Tomasz Zieliński[1], Andrew J. Millar[1*.]

1: Centre for Engineering Biology, School of Biological Sciences, University of Edinburgh, Edinburgh EH9 3BF, UK

2: Institute of Ecology and Evolution, University of Edinburgh, Edinburgh EH8 9AA, UK

*: Corresponding Author

HD, ORCID 0000-0002-7044-7321, hdeeb@ed.ac.uk

HYW, ORCID 0009-0004-9503-9705, hweeyunwong18@gmail.com

TU, trishausman174@gmail.com

MK, ORCID 0000-0002-8696-6978, mkutzer@ed.ac.uk

TZ, ORCID 0000-0002-0194-5706, tomasz.zielinski@ed.ac.uk

AJM, ORCID 0000-0003-1756-3654, andrew.millar@ed.ac.uk

Abstract:

The evolution of research practices emphasizes the role of Open Data in fostering transparency and collaboration. This study evaluates the adoption of Open Data practices in the School of Biological Sciences at a research-intensive university in the United Kingdom. Our analysis of research data sharing from 2014 to 2022 manually reviewed 193 journal articles against criteria for Openness and FAIRness, including the Completeness of data shared relative to data generated. The findings reveal an uptick in data completeness and reusability, with a clear influence of data type, and genomic data being shared more frequently than image data. Data availability statements (DAS) and preprint sharing show a strong correlation with higher Openness and FAIRness scores. Since 2016, when the FAIR guidelines were published, data Reusability increased along with the inclusion of Data Availability Statements. On the other hand, since the COVID-19 pandemic, we have found a substantial increase in preprint sharing and significant improvements in Completeness, Reusability, and Accessibility scores. This paper documents a local research institute's journey towards Open Data, addressing the changes and advocating for best practices to nurture this progression.

Introduction

In today's scientific ecosystem, there has been a growing emphasis on the principles of “Open Science”. This movement encompasses a broad spectrum of practices aiming to make research more transparent, reproducible, and universally accessible (1). Central to Open Science are the objectives of fostering global collaboration, accelerating research progress, and ensuring that research findings are comprehensible not just to the scientific community, but also to the broader public. Yet, merely making data open does not ensure its reusability, especially if encumbered by restrictive licenses (2). Moreover, where reusable public data exists, this alone does not guarantee the reproducibility of the associated results (3).

To address these nuances, the concept of FAIR data – emphasizing data that is ‘Findable’, ‘Accessible’, ‘Interoperable’, and ‘Reusable’ – has been introduced (4), to maintain the integrity of shared research data. Tailoring the FAIR guidelines specifically for each research area can significantly improve how data is managed and used in these domains (4). The ultimate vision, as articulated by UNESCO, is to make data as open as possible, yet keep it restricted when deemed essential, especially in contexts involving sensitive information (5). Reaching this balance, however, is complex, notably when data holds commercial value or pertains to proprietary interests (6).

Biomedical research underscores the balancing act between open access and data protection. Intellectual property rights and considerations surrounding human data disclosure pose consistent challenges (7,8). Though data anonymization has been advanced as a solution, risks of re-identification remain, especially among vulnerable populations (9,10). Additionally, the sharing of anonymized data may not always be covered by the original consent agreements, adding a layer of complexity to privacy protection. However, Open sharing's pivotal role was evident in the rapid COVID-19 vaccine development (11,12) and holds significant potential for addressing socio-economic challenges in non-communicable diseases research (13). The broader adoption of open science faces hurdles. The scientific community's declared support for open sharing does not always translate to practice (14,15). Concerns about dual-use research data and their misuse (16) and potential errors discourage some researchers (14).

Probably the biggest challenge is that effective data sharing requires more than openness; it involves significant labor and dedication to ensure quality and compliance. Therefore, current initiatives by various stakeholders underscore the momentum toward ensuring data openness and FAIRness (17), but also involve complex tasks such as standardizing data formats, ensuring data security, and facilitating data interoperability. Therefore, gaps between data availability claims and actual accessibility are evident (18,19). This discrepancy becomes more apparent when considering the methodologies of past studies (9,20), which predominantly focused on assessing the quality of the data shared, rather than measuring the proportion of available data that was effectively shared

Reflecting these evolving standards, the University of Edinburgh updated its Research Data Management Policy in 2021. Initially, in 2011 (21), the policy focused on data integrity, confidentiality, and retention, with Principal Investigators taking the lead in data management. By 2021 (22), it had evolved to include adherence to FAIR principles and GDPR compliance, highlighting the importance of data accessibility, interoperability, privacy, and security. This progression mirrors the broader movement towards more transparent, accessible, and collaborative research practices.

Understanding our partly open science landscape is now vital, in order to address the remaining barriers. Therefore, our study aims to evaluate both the current state and the progress of Open Data Practices in the School of Biological Sciences at the University of Edinburgh. This involves assessing the impact of significant events, such as the introduction of FAIR principles and the COVID-19 pandemic, on data sharing trends. Our methodology includes a specialised scoring system to evaluate FAIRness and Openness in scientific publications. By comparing practices across different research departments, we intend to provide valuable insights for academia and policymakers alike. The goal is to offer clear, actionable recommendations that promote enhanced transparency and collaboration in scientific research.

Methods:

Definition of Data:

The term "Data" refers to both qualitative and quantitative information necessary for reproducing research results. This includes primary data, such as raw experimental results; secondary data, like derived analytical values, and the software code used to derive those values. Data also includes images and videos used for conducting analyses, but those employed solely for illustrative purposes are excluded from this definition. When quantitative analysis is performed on images, these images are considered raw data, and the resulting numerical values are considered processed data.

Sampling Framework and Selection Process

We selected research groups at the University of Edinburgh, specialising in Non-Communicable Diseases (NCDs) and Genetics or in Infectious Disease topics like Immunology and Vaccines, Molecular Pathogenesis, and Disease Dynamics.

Each group needed a substantial publication record, with at least 10 journal articles published between 2014 and 2022. From each selected group, we randomly chose journal articles for each year within our study period to ensure a representative sample. Our focus was strictly on original research papers that generated new datasets, excluding other publication types such as reviews, editorials, commentaries, and clinical trials, which did not align with our data assessment criteria. This selection process was carried out by two biology students, who manually reviewed and verified each article. The selected articles were then cataloged on an Excel sheet, with each research piece categorised under either Non-Communicable Diseases (NCD) or Infectious Diseases (InfD).

Scoring Criteria:

The scoring system was developed based on those established by Roche et al. (9) and the FAIR Data Self-assessment tool by the Australian Research Data Commons (23). We tailored the scoring criteria, pertinent to our study's objectives. A key adaptation involved refining the scoring scale from five levels, as used by Roche et al., to four. This reduction was motivated by a need to distinctly address file format and metadata quality without conflating them with raw versus processed data considerations, thereby eliminating potential overlap and ambiguities in evaluations.

While Roche et al. (9) sampled from a popular data repository Dryad, where sampling started from archived datasets, this study involved selecting articles while being blinded as to whether any data was shared. When articles reused previously published data, a similar approach to Hamilton et al. (20) was taken. The four criteria – 'Completeness', 'Reusability', 'Accessibility', and 'Licence' – were established to assess independent aspects of data sharing.

1- **Data Completeness:** To evaluate completeness, we assigned scores from 1 to 4, with 4 indicating complete data sharing necessary for reproducing all analyses and results. The highest score did require sharing processed data at each analysis step, provided raw data and sufficient description of data processing steps were available. Scores of 2 and 3 indicated minor and major data omissions, respectively. When a score of 1 was assigned, indicating no data sharing, the assessment of the subsequent criteria was consequently assigned a score of 1 as well (Not Scored). The absence of shared data negated the possibility of evaluating reusability, accessibility, and licence clarity.

Additionally, for each article included in our study, we comprehensively assessed the completeness of data, irrespective of its location - whether in figures, tables, supplementary materials, or external repositories (Table 1).

2- **Data Reusability:** Reusability scores, ranging from 1 to 4, were based on data being in a non-proprietary, human- and machine-readable format, and accompanied by informative metadata. A score of 4 required fulfillment of all these aspects. The criteria for classifying files according to their non-proprietary, human- or machine-readable format are detailed in Table S1 of the supplementary material, adapted from Roche et al (9). The inclusion of metadata in this criterion highlights its indispensable role in preserving the data's long-term utility (Table 1). Proprietary formats were penalized unless they were community-accepted or mandated by public repositories (24,25).

3- **Data Accessibility:** Accessibility was assessed on whether the data was accessible, had a PID, and was archived in a public repository. Scores ranged from 1 to 4, with 4 indicating fulfillment of all conditions. Data with a PID in a public repository but inaccessible or requiring permission received the lowest score, reflecting our focus on data openness. Scoring details are in Table 1.

4- **Data Licence:** The licence type was scored from 1 to 4. Open licences received the highest score, restrictive licences scored 3, and data without a specified licence or governed by the article's licence scored 2. Criteria are detailed in Table 1.

Additional Data Evaluation Criteria:

Our study incorporated several supplementary variables, including the examination of specific data types shared. Image data was categorized as Not Applicable (NA) if image analysis was not pertinent to the study, Not Shared (0) if image analysis was conducted without data sharing, and Shared (1) if image data were both analyzed and made available. These evaluations also encompassed unprocessed images suitable for re-analysis. Similarly, genomic data sharing was assessed (comprising nucleotide sequence-based assays including transcriptomics), with NA indicating no genomic analysis was performed, 0 for genomic data not archived or incompletely shared, and 1 for fully archived genomic data. Additionally, we tracked the publication of preprints, utilizing resources like Google Scholar and the bioRxiv preprint server for verification.

Data Availability Statements (DAS) in each article were only recognized if explicitly titled as such or with similar designations. DAS were categorized as 'Shared' when indicating unrestricted data access, or 'Not Shared' when access was conditional. Articles without a specific DAS section were marked as 'Not Presented`. Finally, the availability of any code or analytical procedure used in the research was documented, irrespective of where they were mentioned within the publication.

Table 1: . Scoring Criteria used for the Assessment of Articles

Data Completeness		
Score	Description	Criteria
4	Exemplary	All the data necessary to reproduce the analyses and results (in practice) are present within the article (e.g. as supplementary information, figures)

		or archived in external repositories. Both raw and processed datasets from all methods utilised and mentioned in the study are provided.
3	Good	Most of the data necessary to reproduce the analyses and results (in practice) are present within the article (e.g. as supplementary information, figures) or are archived in external repositories. Processed data from all methods utilised and mentioned in the article are provided, lacking only a small amount of raw datasets.
2	Large Omission	Main analyses in the paper cannot be redone because essential datasets are missing AND/OR only summary statistics (e.g. means, standard deviation) obtained from methods utilised and mentioned in the article are archived, no raw data provided.
1	Poor	Neither processed nor raw data are archived/present in the article OR the wrong data are archived.

Reusability		
Score	Description	Criteria (For best dataset)
4	Exemplary	Data is archived in a (1) non-proprietary, (2) human- and machine-readable file format that facilitates data aggregation and can be processed with both free and proprietary software (e.g., csv, text, stored in a specialist repository that requires a reusable format). (3) Highly informative metadata (such that column headings, abbreviations, and units can be understood in isolation from the original paper).
3	Good	Data is archived in (1) a non-proprietary OR (2) a human- and machine-readable file format that facilitates data aggregation and can be processed with both free and proprietary software. AND (3) Metadata must at least be sufficiently informative to be understood when combined with the paper.
2	Average	Data is archived in (1) a proprietary, (2) human- but not machine-readable file format (e.g., pdf, jpeg). AND (3) Metadata is not sufficiently informative when combined with the paper.
1	Poor	Not Scored

Accessibility		
Score	Description	Criteria (For best dataset)

4	Exemplary	Data is accessible, has a Persistent Identifier (PID) assigned and stored in an online public repository (e.g., Figshare).
3	Good	Data is accessible, has either a Persistent Identifier (PID) assigned OR stored in an online public repository.
2	Average	Data is accessible but does not have a Persistent Identifier (PID) assigned and is not stored in an online public repository (e.g. data shared in the supplementary information section of the article).
1	Poor	Not Scored

Licence		
Score	Description	Criteria (For best dataset)
4	Exemplary	Data has a permissive licence (e.g. CC0, CC-BY) and has an Open Software licence (if applicable).
3	Good	Data has a restrictive licence (e.g. CC-BY-SA/NC) and has a Closed Software licence (if applicable).
2	Average	No licence is provided for data or software.
1	Poor	Not Scored

Data Extraction and Quality Control Assessment

The collected data, including both scoring assessments and supplementary variables, were cataloged in an Excel spreadsheet. Researchers were assigned to extract and evaluate data specifically from either non-communicable disease (NCD) or infectious disease (Inf) studies. To ensure the reliability of our assessment, a quality control step was incorporated. A third researcher independently reviewed a random selection of 20% of the papers from each evaluator's dataset, covering the entire spectrum of scoring values. This blind reassessment aimed to verify the consistency and objectivity of the initial evaluations. The congruence in scoring assessments between the original and secondary evaluations ranged between 86% and 90%. Consequently, the initial assessments were retained for final analysis.

Statistical Analysis

The descriptive statistics were presented as frequency and percentage. Chi-squared tests determined the variance between the two research fields and the changes in Data Availability Statements (DAS) and preprint publications before and after the implementation of FAIR principles and the onset of the COVID-19 pandemic. The Mann-Whitney U test was utilised to evaluate the distribution differences in scoring criteria between the Inf and NCD papers. An ordinal regression model was used to test for changes over time in the four scoring

criteria from 2014 to 2022. Additionally, ordinal regression analysis measured the influence of sharing variables such DAS, and preprint status on the scoring criteria. Significance was set at $p < 0.05$, and results were expressed as odds ratios with corresponding 95% confidence intervals. Additionally, thorough checks were conducted to validate the assumptions underlying each statistical model used in this analysis. Details of these assumption checks, along with the complete analytical methodology, have been documented and are available on GitHub repository link mentioned in the Data Availability Statement. All the analyses were performed using R and RStudio Software (Version 4.2.2).

Results:

Study sample overview:

From 2014 to 2022, the School of Biological Sciences at the University of Edinburgh published 3,492 research papers. Our study analysed a sample of 193 papers, representing 5.53% of the total publications in this period. This sample was divided based on the primary research focus, with 111 papers (58%) focused on Non-Communicable Diseases (NCD) research and 82 papers (42%) centered on Infectious Diseases (InfD) research. A sizable portion of the studies in our sample, totaling 145 papers, were published following the dissemination of the FAIR principles in 2016 (Table S2: Supplementary Material). Post the onset of the COVID-19 pandemic in 2020, the number of published papers in our sample was 52.

We also observed varying levels of openness in research practices. Regarding Data Availability Statement (DAS), 37.31% of the total publications included a DAS with Shared Data. In contrast, 15% offered data upon request, thus considered as "Not Shared", and a significant 47% did not present a DAS. Additionally, there was a high percentage of the papers (99%) shared their used analysis program, and a notable majority, 48 out of 69, actively shared their code when used. This high percentage of code-sharing indicates a strong tendency towards openness in computational research. However, our findings show limited adoption of preprints, with only 22% of papers having associated preprints (Table S2: supplementary material). Across these different aspects – DAS presence, code sharing, and preprint utilisation – our analysis indicated similar patterns in data management and dissemination across the research domains. (Figure S1: Supplementary Material).

Differences in Data-Sharing Practices by Data Types

In our analysis, we identified notable differences in data-sharing practices across different types of data. 116 papers ustilised image data, yet only 11 of these have shared their image data. This contrasts with genomic data: out of 103 papers that used genomic data, a sustainable number, 71 papers, shared their genomic data. This significant discrepancy – with only about 9.5% of image-data-using papers sharing their data, compared to 69% for genomic data – underscores a marked preference in the scientific community for sharing genomic data over image data. This trend is observed consistently across studies, whether they focus on Non-Communicable or Infectious Diseases.

Assessment of FAIRness and Openness Criteria in Research Publications:

In the Completeness criterion (Figure 1), only 12% of the publications shared all the available data and scored 4. Most articles (66%) achieved scores of 2 or 3 indicative of moderate completeness and 22% of the publications had a score of 1, where no data were shared. The 42 articles with Completeness score 1 were also scored 1 (Not Scored) for the other criteria. For Reusability and Accessibility, the largest segment of publications fell into the highest score range, scoring 4, with 41%, and 50% respectively. Licence clarity showed the most favourable results, with a substantial 68% of publications having a permissive licence for their shared data, with an open software licence (Figure 1). Where bioscience data were shared, they were often shared competently but only 32% of articles shared most or all their data (Completeness score of 3 or 4).

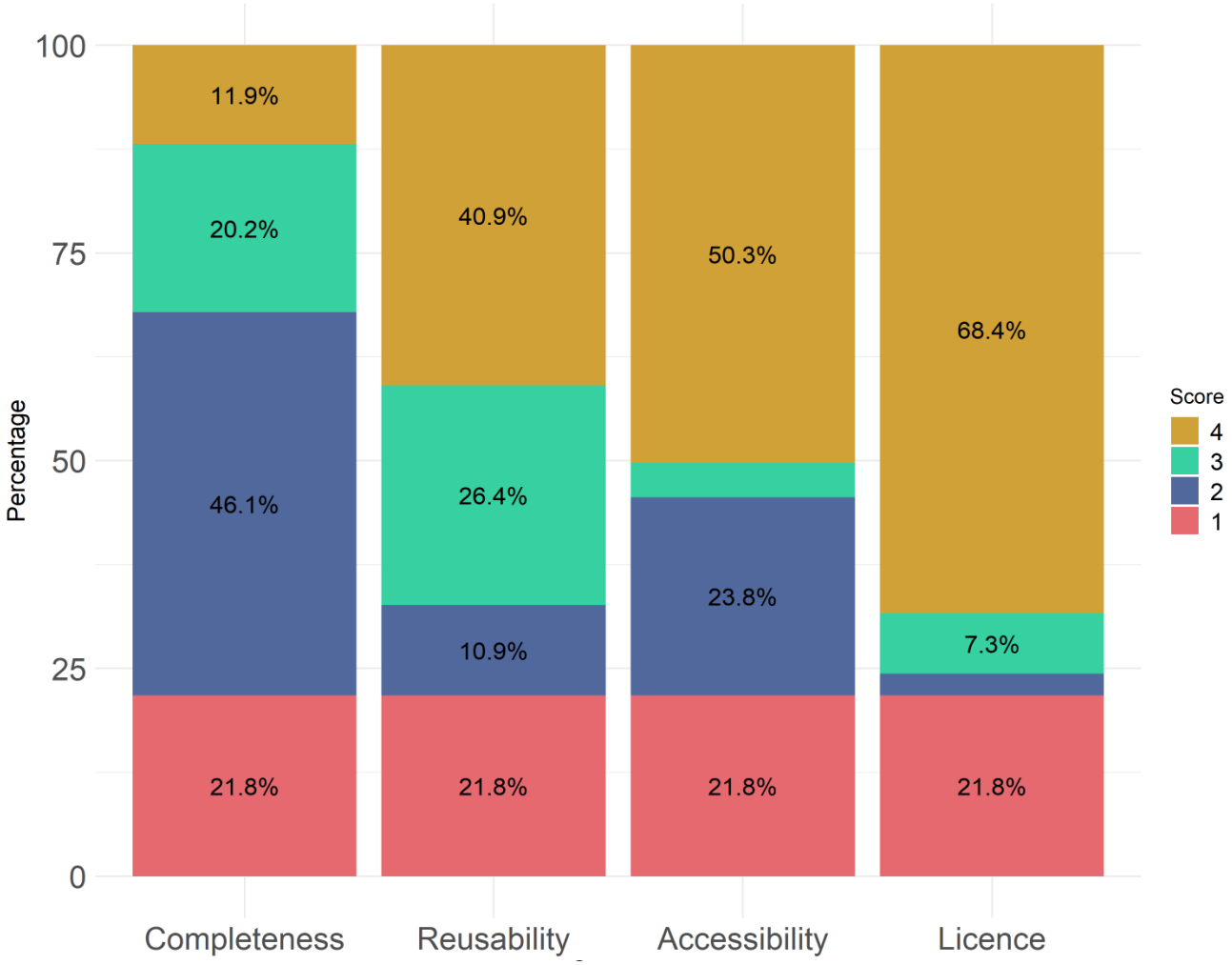

Figure 1: Scores of the four Criteria (Completeness, Reusability, Accessibility and Licence) in all the research publications (n=193).

Ordinal Regression Analysis: Changes in Data Sharing over Time

Plotting the scores over time (Figure 2) suggests a gradual improvement in Accessibility and Licence, with a higher proportion of papers receiving the highest score by 2022. Completeness and Reusability scores seem to show more variability over the 9 years. Nonetheless, comparing the first and last years (2014 to 2022) suggests a striking increase in Completeness: only 10% of the 2014 articles shared most or all of their data, scoring 3 or 4. By 2022, that fraction had risen to 63%. We used an ordinal regression model to test the significance of these changes across the entire set of articles.

The ordinal logistic regression analysis revealed a statistically significant year-over-year increase in the odds of higher scores across all evaluated criteria, suggesting a positive temporal trend in data quality measures. The Completeness criterion showed a modest but consistent improvement (OR = 1.19, 95% CI [1.08, 1.33]), indicating a 19% increase in the odds of a higher score each year. While the Reusability criterion demonstrated a stronger positive trend (OR = 1.24, 95% CI [1.12, 1.38]). Similarly, there has been a significant

and consistent increase in Accessibility scores, indicating a 21% annual increase in the odds of achieving higher scores in this criterion, alongside Licence criteria (OR = 1.16, 95% CI [1.03, 1.31]) (Table S3 Supplementary material). Figure 3 provides a visual representation of the changes in scoring criteria from 2014 to 2022 detected by the ordinal regression model. Each criterion exhibits a distinct trend, with an overall increase in the likelihood of higher scores, particularly for Reusability and Licence criteria, suggesting improvements in these areas over the observed period.

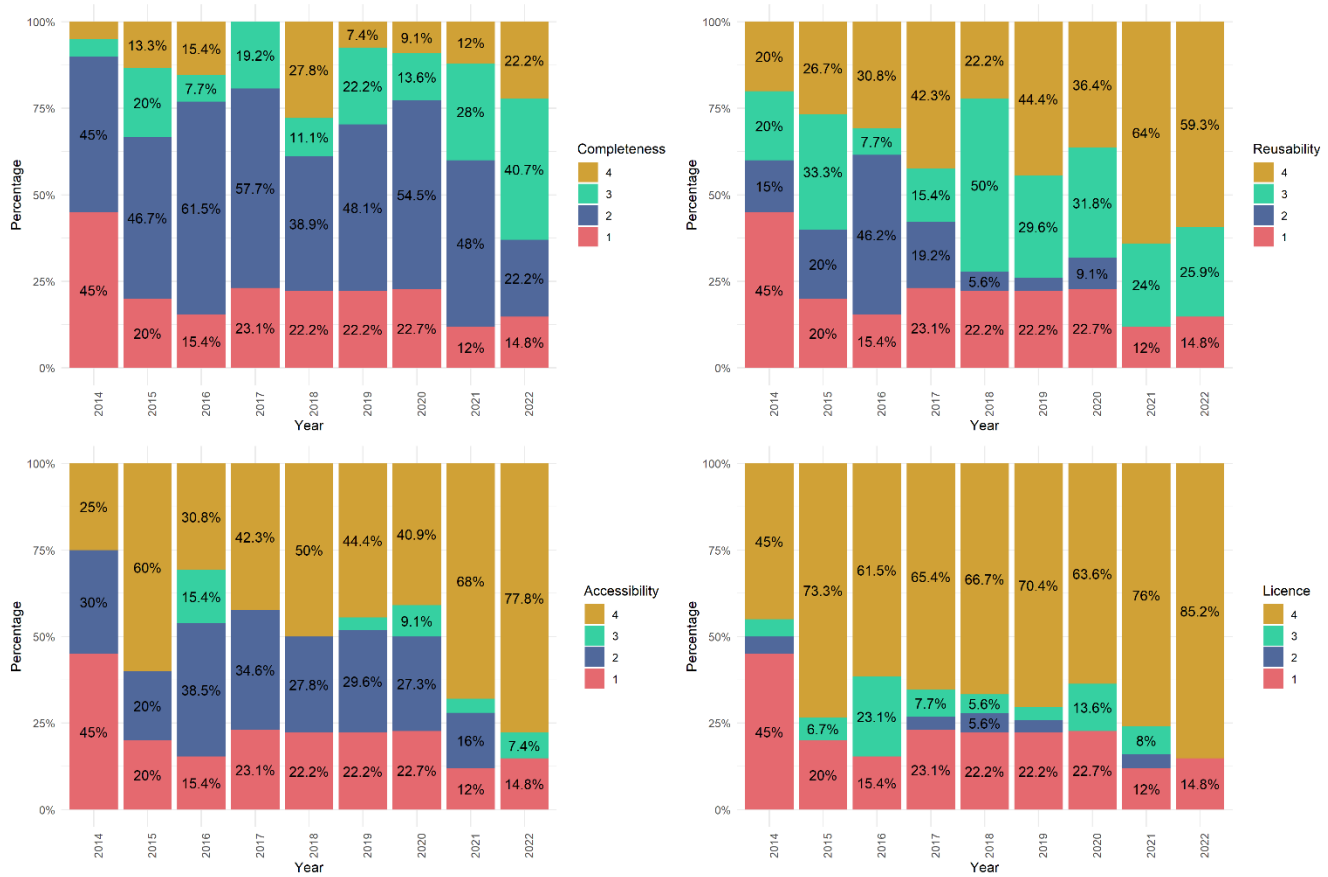

Figure 2: The distribution of scoring criteria over nine years, reflecting changes in Completeness, Reusability, Accessibility, and Licence of 13- 27 research publications annually.

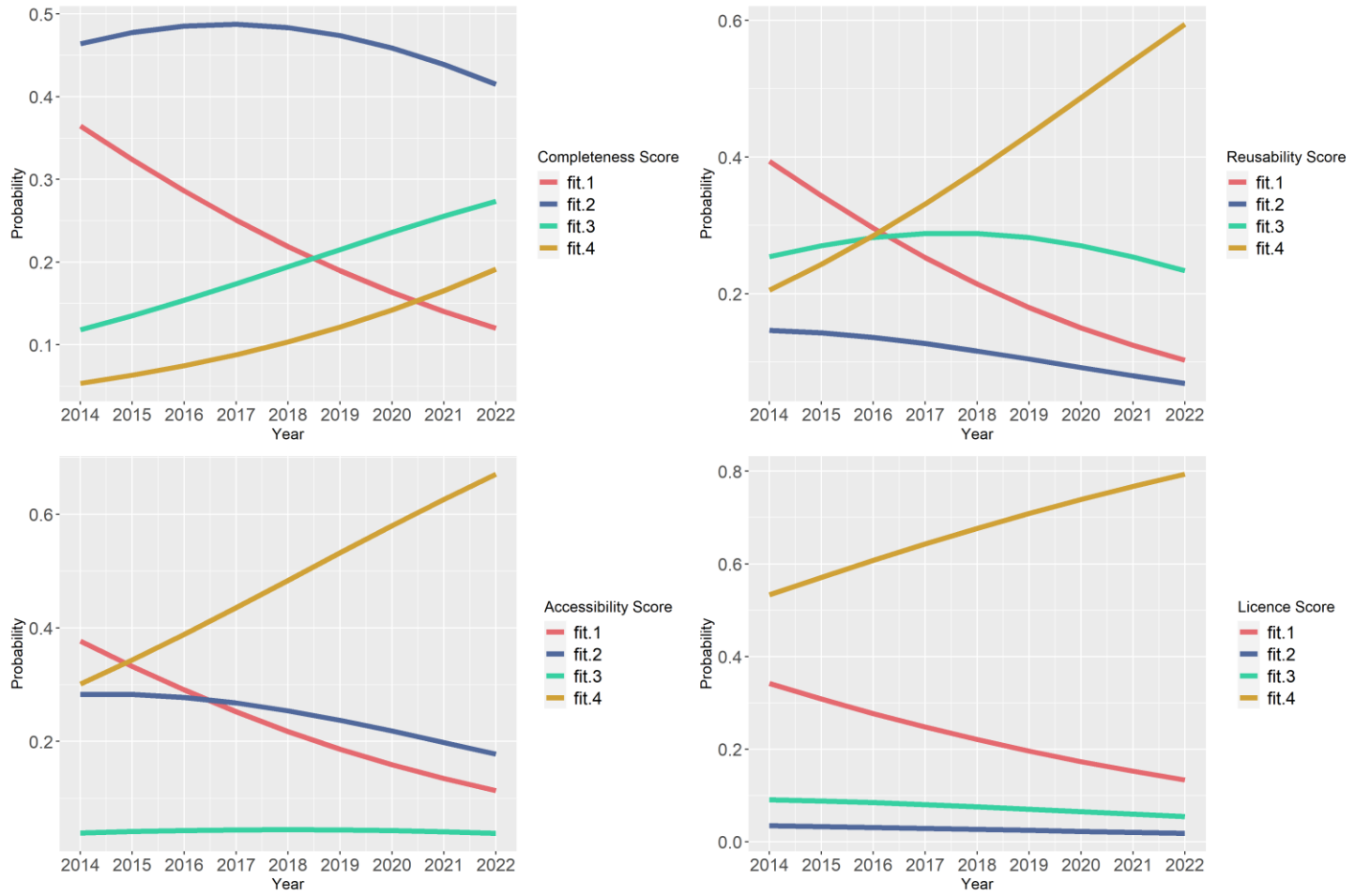

Figure 3: Evaluation of the changes in the four Scoring Criteria for 193 bioscience papers from 2014 to 2022: Ordinal Regression model.

Ordinal Regression Analysis: Influence of Data sharing Variables: DAS, and Preprint Status on Scoring Criteria

We then tested whether high-quality data sharing was related to characteristics of the articles other than their publication date, namely whether a preprint was shared before publication, and the presence or strength of a Data Availability Statement (DAS). Articles with a DAS that signals Shared data had markedly higher Completeness scores (OR = 13.16, 95% CI: 6.29 - 27.51, $p < 0.001$) (Table 2). This significant pattern also held for the Reusability, Accessibility, and Licence criteria, underscoring the positive relationship of journals' open data policies with data sharing. Even papers where data were only signalled as accessible upon request scored higher than those without a DAS, albeit with slightly lower odds ratios (Table 2). Papers with preprints were not strongly associated with better data sharing, demonstrating a positive correlation only with Completeness (OR = 2.23, 95% CI: 1.14 - 4.40, $p < 0.05$). Articles with a strong DAS were much more clearly associated with high-quality data sharing in our sample, compared to articles with preprints. However, we also found striking exceptions to this pattern (see DAS Status Relative to Completeness).

Sharing Variables		Completenes	Reusability	Accessibility	Licence
Data Availability Statement (DAS)	DAS with Open Data	13.16 (6.29 - 27.51)***	6.75(3.51 - 12.98)***	5.86 (3.02 - 11.36)***	6.40 (2.80 - 14.60)***
	DAS with Data upon request	3.26 (1.39 - 7.63)**	5.52 (2.36 - 12.92)***	5.02 (2.05 - 12.29)***	5.31 (1.69 - 16.69)***
	Not Presented	Baseline Group			
Preprint	Yes	2.23 (1.14- 4.40)*	1.08 (0.55- 2.13)	1.66(0.79 - 3.49)	2.38 (0.81 - 6.85)
	No	Baseline Group			

Table 2: Ordinal Regression Analysis Examining the Impact of Data Availability Statement (DAS), and Preprint Status on Scoring Criteria

OR interpretation: OR values reflect the multiplicative increase in the odds of achieving higher criterion scores, where each unit of DAS categories or the presence of a preprint can elevate the odds by up to several times, depending on OR of the specific criterion assessed.

Note:

- The numbers presented are Odds Ratios (OR) with 95% Confidence Intervals (CI).
- Statistical significance is indicated by asterisks:
  *p-value < 0.05, **p-value < 0.01, and ***p-value < 0.001.

Before-and-After Analysis: Scoring Criteria Changes Surrounding FAIR Implementation (2016) and COVID-19 Pandemic (2020)

Comparing the years before and after these key events revealed some notable changes, which we formalized with the Mann-Whitney U tests. The analysis of the 145 papers published after the adoption of FAIR principles in 2016 revealed a marked improvement in 'Reusability' scores compared to the 48 papers from before 2016, with a significant P-Value of 0.002. This indicates an enhanced focus on data reusability in research publications, with better file formats and metadata. The period following the onset of the COVID-19 pandemic, particularly the years 2021 and 2022, marked a significant shift with a notable improvement in data Completeness, evidenced by a P-Value of 0.001. This trend was accompanied by remarkable enhancements in 'Reusability' and 'Accessibility' scores. Specifically, 'Reusability' scores showed significant improvements across various paper types: overall ($p < 0.001$), non-communicable diseases (NCD) papers ($p = 0.006$), and infectious diseases (InfD) papers ($p = 0.039$). Similarly, 'Accessibility' scores followed this upward trend, with overall papers ($p < 0.001$), NCD ($p = 0.002$), and InfD ($p = 0.037$) all showing substantial gains. These improvements suggest a pronounced shift towards better data sharing and accessibility practices in response to the pandemic's challenges. Notably, this analysis includes 52 papers published after the pandemic's onset in 2020, compared to 141 papers from the pre-pandemic era. However, the 'Licence'

criterion did not show significant changes during this period across all types of papers, indicating areas where further progress is needed.

Before-and-After Analysis: Changes in Sharing Variables (DAS, Preprints) Following FAIR Implementation (2016) and During the COVID-19 Pandemic (2020)

Our analysis has identified a clear trend of increasing frequency in the issuance of data availability statements (DAS), coupled with a notable enhancement in the dissemination of complete datasets and a marked increase in sharing preprints (Figure S2 and Figure S3: Supplementary Material).

However, the dissemination of the FAIR data principles in 2016 coincided with a change in the landscape of research data sharing. In 2014-2015, only 16.67% of studies signalled Shared data with a data availability statement (DAS), which increased to 46.21% post-FAIR (2017-2022). This substantial rise highlights a growing commitment to data transparency in the research community. Conversely, the percentage of studies without a Data Availability Statement saw a notable decrease (Table S4: Supplementary Material).

The COVID-19 pandemic in 2020 profoundly impacted the sharing of preprints and data availability in our sample. There was a five-fold increase in the sharing of preprints, from a mere 11.35% before 2020 to 53.85% in 2021-2022 (Table S5: Supplementary Material). The trend towards open data sharing also continued, with 55.77% of studies in 2021-2022 having a DAS signalling Shared data, compared to 32.62% before 2020.

Do Data Availability Statements predict data sharing?

Figure 4 demonstrates a clear correlation between data sharing, as indicated by completeness scores, and the status of the Data Availability Statement (DAS) in these research publications, showing the details that underlie the ordinal regressions (Table 2). Higher DAS status is observed in papers with elevated completeness scores, so 22 of the 23 papers that shared all their data (Completeness 4) signalled this with a 'Shared' DAS. Indeed, the DAS scored as 'Shared' often specifies that all data are shared. Similarly, as the completeness score decreases, the likelihood of a paper lacking any DAS increases: 33 of the 42 papers that shared no data also lacked a DAS. However, the converse relationship is much weaker. Eight articles that indicated data were available upon request (a 'Not Shared' DAS) had shared most of their data (Completeness 3), and articles with a 'Shared' DAS were represented in all Completeness categories. At the extreme, six papers that claimed all data were available in their DAS paradoxically received a Completeness score of 1, indicating that no data were shared. This discrepancy highlights a gap between the declared availability of data and the actual sharing of complete datasets.

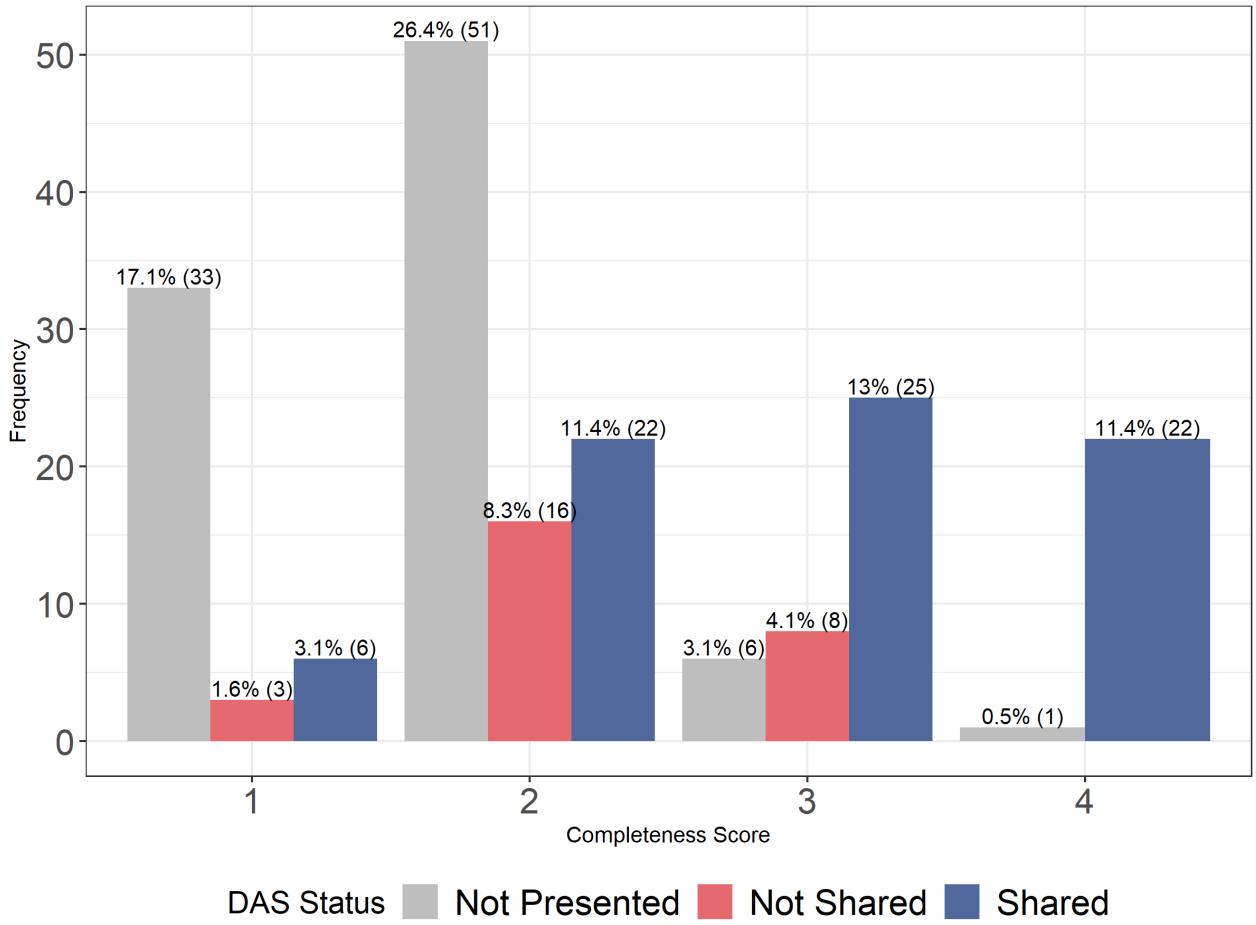

Figure 4: the Distribution of the Data Availability Statement (DAS) status across Completeness score categories for the 193 research papers.

Discussion:

Our analysis of 193 bioscience research papers from the University of Edinburgh has shown a noteworthy trend: a progressive shift towards sharing more research data and sharing it more FAIRly. This trend was clearly more noticeable after COVID-19. The challenges of the COVID-19 pandemic marked a paradigm shift in research transparency. The increased incorporation of Data Availability Statements in these papers is a testament to this shift, often indicating a higher level of data completeness and reflecting a deeper commitment to open research practices. These patterns are consistent with the ordinal regression findings, underscoring a movement towards greater transparency in research practices over recent years.

Our manual scoring of the openness and FAIRness of data from published papers laid the groundwork for a detailed evaluation, starting with an initial assessment of whether the data was shared (completeness score) before delving into the specifics of how and where it was disseminated. Similar approach was previously applied in the cancer field (20), and is more comprehensive than studies starting evaluation from already shared datasets(9).

This manual scoring process revealed a striking difference in the sharing practices across various data types, 73 papers shared genomic data, whereas only 11 shared imaging data, underscoring a notable disparity. This disparity can be largely attributed to the historical development and availability of specialised data repositories, which is related to the scale and complexity of the data. Genomic data (sequence-based assays including transcriptomics) has benefited from established, long-running databases like GenBank and GEO, fostering a culture of sharing that is well-entrenched within the scientific community (26,27). Moreover, most publishers require the deposition of genomic data into designated repositories at paper submission. This contrasts with imaging data, where raw microscopy images were rarely shared. Part of the problem is the lack of a suitable repository; the Image Data Resource (IDR) and BioImage Archive (BIA) are relatively new and have only recently begun to accept substantial contributions (28). Alternatively, a general, data-agnostic repository like Figshare or Zenodo could be used, but they do not offer microscopy image viewers and the upload of large datasets can be problematic. Microscopy images with multistack datasets are often more than 10GB in size and their transfer with website wizards is unreliable. The cost of storing imaging data generally can also present financial challenges (29). As initiatives like BIA are in their infancy, the practice of sharing imaging data is not as widespread or normalised as in genomics. However, the increasing emphasis on data sharing across the board, driven by journals and funders' mandates, might soon lead to more balanced data-sharing practices across different data types (30).

The incremental improvement in the adoption of open research practices within the University of Edinburgh's biosciences sector, particularly in terms of the completeness of data sharing, aligns with findings from previous research (9,15); one study assessed datasets in the digital repository Dryad (9), while another in the US collected insights through surveys and interviews (15). Despite these improvements, the pace at which these practices have been embraced remains relatively slow, with only 23 out of 193 papers fully sharing all their data. This sluggish progression can be attributed to several factors (14). A primary reason might be the recent overhaul of the University's 'Research Data Management' policy. Before its update in 2021, this policy had remained unchanged since 2011 and only recently began to integrate the FAIR principles (22). This policy renewal could have led to a lag in the full implementation and widespread acceptance of open science practices. Additionally, the stance of major funding bodies, such as the Wellcome Trust, on open research data worths attention. In 2016, the Wellcome Trust, along with other leading research funders, endorsed the

UK Concordat on Open Research Data (31). This concordat advocates for maximizing data sharing and acknowledges the need to include provisions for data management and sharing costs within research grants. The interplay between university regulations and funder mandates plays a critical role in shaping open research practices, suggesting that the alignment and evolution of these policies are key drivers in the adoption and enhancement of open science within the academic community. It's also noteworthy that 'Completeness' predominantly evaluates open and unrestricted data sharing. However, some data, despite being shared following FAIR principles, may reside on platforms or repositories requiring registration or membership, thus introducing a layer of conditional access to what might otherwise be considered openly shared data.

The adoption of FAIR guidelines, emphasizing the reusability of datasets, has likely influenced researchers at the University of Edinburgh to adopt more open file formats and provide comprehensive metadata, with 130/193 has achieved the highest Reusability score (3-4). This improvement is aligning with the increasing demands of funders and organizations (32,33). However, a notable observation in this context is the differential impact of the COVID-19 pandemic on various research areas. While the exact reasons remain speculative, it is plausible that the critical role of data sharing during the pandemic encouraged many bioscientists to share their data (34). In our sample, this change was more pronounced among research groups studying Non-Communicable diseases rather than Infectious disease. Another explanation is that during the lockdowns, bench researchers had an opportunity to learn new skills like data sharing and had more time to prepare datasets for public sharing while laboratory access was limited the post-COVID period has witnessed substantial advancements in both FAIRification and open data practices. the surge in COVID-19-related preprints and their subsequent influence on research dissemination practices is notable. The rise in non-COVID preprints post-pandemic (35) mirrors this trend and is consistent with our findings, indicating a broader, potentially transformative impact across various scientific fields. This shift potentially signifies either a temporary adaptation in response to the global health crisis or a more profound and lasting transformation in the culture of scientific research.

The introduction of open data policies in scientific publishing, while a positive step, has not led to a comprehensive adherence among researchers, particularly in the aspect of completeness of datasets. This issue is highlighted by the presence of Data Availability Statements (DAS) in publications, which, despite their intent, are not accompanied by the complete sharing of data necessary for reproducing research results (36). This contradiction was seen in our results, where 6 papers with a DAS that signalled “all data were available” but had no data accompanied the publications in any form. While journals like those mentioned by Bloom et al. (37) and Rebecca and Iain (38) have mandated the inclusion of DAS, there remains a discrepancy between the claim of "all data available" and the actual data shared. This might have originated in some authors' confusion between a visual display of the data, such as a graph that is published, and the numerical values of the data that should accompany it. Whatever its origin, this gap in compliance persists due to the lack of enforcement mechanisms within the publishing process. The responsibility often falls on editors and peer reviewers, who may lack the resources or the expertise to thoroughly verify the completeness and transparency of shared data (39). This issue was highlighted by the UK House of Commons’ Science, Innovation and Technology Committee in 2023 (40).

While this study reflects the status of open data within the University of Edinburgh's biosciences department, it does encounter limitations. Its focus on a single institution may not reflect broader trends in open data practices across different research communities or disciplines. Additionally, the quantitative approach

predominantly employed does not capture the nuanced motivations and challenges researchers face in data sharing (41). The study's timeframe also restricts the ability to fully assess the long-term impact of recent policy changes and global events, such as the COVID-19 pandemic, on open research practices.

Considering the findings from this study, funders and research institutions need to prioritise enhancing data repository infrastructures, particularly for underrepresented data types like imaging data. Simultaneously, stronger enforcement of open data policies should be implemented, through improved training and standardized guidelines for editors and peer reviewers. Additionally, fostering a culture that values open data practices is crucial, which could be achieved through educational initiatives and awareness programs targeted at researchers. These combined efforts are essential for advancing the quality, transparency, and impact of scientific research in the era of open data and collaborative discovery.

Standpoint

The authors worked at the School of Biological Sciences, University of Edinburgh during the study, in the Biological Research Data Management team (BioRDM). Our team grew from the need for data management in our interdisciplinary biology research projects in the Centre for Systems Biology at Edinburgh (later SynthSys, now Centre for Engineering Biology), since 2008. AJM also chaired the UKRI-BBSRC Review of Data-Intensive Bioscience (2020).

Supplementary Materials

The supplementary materials for this paper, provided as a PDF file, can be accessed through the paper's GitHub repository at https://github.com/BioRDM/InsightsOfOpenPracticesInBiosciences.

Data and Code Availability Statement

All data utilized in this study, along with the detailed code, are available in the GitHub repository dedicated to this paper: https://github.com/BioRDM/InsightsOfOpenPracticesInBiosciences.

Authors' Contributions

AJM designed the study; HYW and TU adapted scoring criteria and performed primary scoring; HD performed secondary scoring, analysed the data, drafted the manuscript and shared the outputs; MK designed the analysis methods; TZ and AJM edited the manuscript and supervised the work.

Conflict of interest declaration

The authors declare no conflicts of interest.

Funding

This project was funded by MRC grant MR/X009726/1. For the purpose of open access, the authors have applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising from this submission.

Acknowledgements

We gratefully acknowledge the University of Edinburgh Library Research Support, for providing the information systems used to identify our colleagues' publications.

References:

1. Vicente-Saez R, Martinez-Fuentes C. Open Science now: A systematic literature review for an integrated definition. Journal of Business Research. 2018 Jul 1;88:428–36.

2. Carbon S, Champieux R, McMurry JA, Winfree L, Wyatt LR, Haendel MA. An analysis and metric of reusable data licensing practices for biomedical resources. PLOS ONE. 2019 Mar 27;14(3):e0213090.

3. Hardwicke TE, Mathur MB, MacDonald K, Nilsonne G, Banks GC, Kidwell MC, et al. Data availability, reusability, and analytic reproducibility: evaluating the impact of a mandatory open data policy at the journal *Cognition*. R Soc Open Sci. 2018 Aug;5(8):180448.

4. Musen MA, O'Connor MJ, Schultes E, Martínez-Romero M, Hardi J, Graybeal J. Modeling community standards for metadata as templates makes data FAIR. Sci Data. 2022 Nov 12;9(1):696.

5. UNESCO. UNESCO Recommendation on Open Science [Internet]. 2020 Mar [cited 2022 Aug 25]. Available from: https://en.unesco.org/science-sustainable-future/open-science/recommendation

6. Lezaun J, Montgomery CM. The Pharmaceutical Commons: Sharing and Exclusion in Global Health Drug Development. Science, Technology, & Human Values. 2015;40(1):3–29.

7. Ballantyne A, Stewart C. Big Data and Public-Private Partnerships in Healthcare and Research. Asian Bioethics Review. 2019 Sep 1;11(3):315–26.

8. Byrd JB, Greene AC, Venkatesh PD, Jiang Xiaoqian, Greene CS. Responsible, practical genomic data sharing that accelerates research. Nature Reviews Genetics. 2020 Oct;21(10):615–29.

9. Roche DG, Kruuk LEB, Lanfear R, Binning SA. Public Data Archiving in Ecology and Evolution: How Well Are We Doing? PLOS Biology. 2015 Nov 10;13(11):e1002295.

10. Adrian Thorogood. International Data Sharing and Rare Disease: The Importance of Ethics and Patient Involvement. In: Zhan He Wu, editor. Rare Diseases [Internet]. Rijeka: IntechOpen; 2020 [cited 2024 Jan 16]. p. Ch. 15. Available from: https://doi.org/10.5772/intechopen.91237

11. Besançon L, Peiffer-Smadja N, Segalas C, Jiang H, Masuzzo P, Smout C, et al. Open science saves lives: lessons from the COVID-19 pandemic. BMC Medical Research Methodology. 2021 Jun 5;21(1):117.

12. Tse E, Klug D, Todd M. Open science approaches to COVID-19 [version 1; peer review: 2 approved]. F1000Research [Internet]. 2020;9(1043). Available from: https://f1000research.com/articles/9-1043/v1

13. Ndubuisi NE. Noncommunicable Diseases Prevention In Low- and Middle-Income Countries: An Overview of Health in All Policies (HiAP). Inquiry: A Journal of Medical Care Organization, Provision and Financing [Internet]. 2021;58. Available from: https://api.semanticscholar.org/CorpusID:237269504

14. Zečević K, Houghton C, Noone C, Lee H, Matvienko-Sikar K, Toomey E. Exploring factors that influence the practice of Open Science by early career health researchers: a mixed methods study. HRB Open Res. 2020;3:56.

15. Tenopir C, Rice NM, Allard S, Baird L, Borycz J, Christian L, et al. Data sharing, management, use, and reuse: Practices and perceptions of scientists worldwide. Lozano S, editor. PLoS ONE. 2020 Mar 11;15(3):e0229003.

16. Smith JA, Sandbrink JB. Biosecurity in an age of open science. PLOS Biology. 2022 Apr 14;20(4):e3001600.

17. UKRIO. 'Concordat on Open Research Data', UK Research and Innovation (UKRI) [Internet]. 2016. Available from: https://www.ukri.org/wp-content/uploads/2020/10/UKRI-020920-ConcordatonOpenResearchData.pdf

18. Kidwell MC, Lazarević LB, Baranski E, Hardwicke TE, Piechowski S, Falkenberg LS, et al. Badges to Acknowledge Open Practices: A Simple, Low-Cost, Effective Method for Increasing Transparency. PLOS Biology. 2016 May 12;14(5):e1002456.

19. Strcic J, Civljak A, Glozinic T, Pacheco RL, Brkovic T, Puljak L. Open data and data sharing in articles about COVID-19 published in preprint servers medRxiv and bioRxiv. Scientometrics. 2022;127(5):2791–802.

20. Hamilton DG, Page MJ, Finch S, Everitt S, Fidler F. How often do cancer researchers make their data and code available and what factors are associated with sharing? BMC Medicine. 2022 Nov 9;20(1):438.

21. The University of Edinburgh. Policy for Management of Research Data [Internet]. 2011 Apr. Available from: https://era.ed.ac.uk/bitstream/handle/1842/38258/RDM-Policy_CMG-20Apr11.pdf?sequence=1&isAllowed=y

22. The University of Edinburgh. Research Data Management Policy. 2021 Jan.

23. Australia Data Research Commons. FAIR Data Self Assessment Tool [Internet]. 2022 May. Available from: https://ardc.edu.au/resource/fair-data-self-assessment-tool/

24. Ternent T, Csordas A, Qi D, Gómez-Baena G, Beynon RJ, Jones AR, et al. How to submit MS proteomics data to ProteomeXchange via the PRIDE database. PROTEOMICS. 2014 Oct 1;14(20):2233–41.

25. EMBL-EBI. File formats in PRIDE Archive - PRIDE - Proteomics Identification Database. 2023; Available from: https://www.ebi.ac.uk/pride/markdownpage/pridefileformats

26. Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, et al. GenBank. Nucleic Acids Research. 2017 Jan 4;45(D1):D37–42.

27. Perez-Riverol Y, Csordas A, Bai J, Bernal-Llinares M, Hewapathirana S, Kundu DJ, et al. The PRIDE database and related tools and resources in 2019: improving support for quantification data. Nucleic Acids Research. 2019 Jan 8;47(D1):D442–50.

28. Williams E, Moore J, Li SW, Rustici G, Tarkowska A, Chessel A, et al. Image Data Resource: a bioimage data integration and publication platform. Nature Methods. 2017 Aug 1;14(8):775–81.

29. Steinhardt I, Bauer M, Wünsche H, Schimmler S. The connection of open science practices and the methodological approach of researchers. Quality & Quantity. 2023 Aug 1;57(4):3621–36.

30. Kim Y. A Comparative Study of Biological Scientists' Data Sharing Between Genome Sequence Data and Lab Experiment Data. In: Taylor NG, Christian-Lamb C, Martin MH, Nardi B, editors. Information in Contemporary Society. Cham: Springer International Publishing; 2019. p. 3–14.

31. Wellcome. Wellcome signs open data concordat [Internet]. 2016 Jul. Available from: https://wellcome.org/news/wellcome-signs-open-data-concordat

32. Boeckhout M, Zielhuis GA, Bredenoord AL. The FAIR guiding principles for data stewardship: fair enough? European Journal of Human Genetics. 2018 Jul 1;26(7):931–6.

33. Hasnain A, Rebholz-Schuhmann D. Assessing FAIR Data Principles Against the 5-Star Open Data Principles. In: Gangemi A, Gentile AL, Nuzzolese AG, Rudolph S, Maleshkova M, Paulheim H, et al., editors. The Semantic Web: ESWC 2018 Satellite Events. Cham: Springer International Publishing; 2018. p. 469–77.

34. Yoon A, Kim Y. The role of data-reuse experience in biological scientists' data sharing: an empirical analysis. The Electronic Library. 2020 Jan 1;38(1):186–208.

35. Watson C. Rise of the preprint: how rapid data sharing during COVID-19 has changed science forever. Nature Medicine. 2022 Jan 1;28(1):2–5.

36. Tedersoo L, Küngas R, Oras E, Köster K, Eenmaa H, Leijen Ä, et al. Data sharing practices and data availability upon request differ across scientific disciplines. Scientific Data. 2021 Jul 27;8(1):192.

37. Bloom T, Ganley E, Winker M. Data Access for the Open Access Literature: PLOS's Data Policy. PLOS Medicine. 2014 Feb 25;11(2):e1001607.

38. Rebecca Grant, Iain Hrynaszkiewicz. The impact on authors and editors of introducing Data Availability Statements at Nature journals. bioRxiv. 2018 Jan 1;264929.

39. Christian TM, Gooch A, Vision T, Hull E. Journal data policies: Exploring how the understanding of editors and authors corresponds to the policies themselves. PLOS ONE. 2020 Mar 25;15(3):e0230281.

40. Science, Innovation and Technology Committee. Reproducibility and Research Integrity [Internet]. 2023 May. Available from: https://publications.parliament.uk/pa/cm5803/cmselect/cmsctech/101/report.html

41. Attenborough R. Finding virtue in open science? Biological scientists' constructions of openness in historical, advocacy and policy contexts. 2021 Jul 31 [cited 2024 Feb 6]; Available from: https://era.ed.ac.uk/handle/1842/37875