Scientometrics
https://doi.org/10.1007/s11192-025-05424-w

Standing on the shoulder of data: data-driven research boosts scientific innovation

Alex J. Yang[1,2] · Star X. Zhao[2,3] · Sanhong Deng[1]

Received: 16 November 2024 / Accepted: 1 September 2025
© Akadémiai Kiadó Zrt 2025

Abstract
The increasing centrality of data in scientific research prompts important questions regarding its influence on the innovative potential of scientific endeavors. While the value of data in enhancing research rigor is well-recognized, empirical studies investigating its effects on scientific progress remain limited. This paper addresses this gap by analyzing the impact of data-driven research on scientific outcomes, utilizing two recent datasets from the life and social sciences. By applying Coarsened Exact Matching (CEM) to pair data-based papers with non-data-based counterparts, we isolate the effects of data utilization on scientific impact, interdisciplinarity, and disruption. Our findings provide evidence that data-based papers demonstrate significantly higher scientific impact, greater interdisciplinarity, and a stronger potential for disruption compared to their non-data-based counterparts. Additionally, data-based papers receive both higher disruptive and consolidating citations, indicating a dual effect that not only challenges existing paradigms but also strengthens and expands the existing body of knowledge. These results underscore the role of data-driven research in shaping the future of scientific inquiry, bridging traditional research paradigms with emerging, data-centric methodologies.

Keywords Science of science · Data-driven research · Scientific impact · Interdisciplinarity · Disruption · Innovation

Introduction

Science, as a dynamic and transformative endeavor, has long been propelled by the contributions of pioneering scholars, upon whose insights contemporary researchers build (Furman & Stern, 2011; Park et al., 2023). In the modern era, the importance of data

✉ Star X. Zhao
xzhao@fudan.edu.cn

✉ Sanhong Deng
sanhong@nju.edu.cn

[1] School of Information Management, Nanjing University, Nanjing, China

[2] National Institute of Intelligent Evaluation and Governance, Fudan University, Shanghai, China

[3] Institute of Big Data (IBD), Fudan University, Shanghai, China

Published online: 13 October 2025

has ascended to a central role in scientific inquiry, akin to the pivotal function of energy in driving industrial revolutions (Liu et al., 2023a; Zanin et al., 2016). The ubiquity of scientific data across disciplines highlights its indispensability, yet the precise ways in which data influences the innovative capacity of scientific research remain underexplored. While the intrinsic value of scientific data in enhancing the credibility and rigor of research is widely acknowledged (Wang & Barabási, 2021), empirical investigations into its transformative impact on scientific advancement have been limited. This gap in the literature leaves a crucial dimension of scientific progress insufficiently understood, posing an important question: how does the use of existing scientific data shape the innovative potential of research?

Recent policy shifts in funding agencies and academic journals have significantly fostered data sharing, both among researchers and between researchers and the broader public (Hemphill et al., 2024; Yin et al., 2022). Such initiatives enhance the transparency of research processes, facilitating the replication, evaluation, and validation of scientific results. In light of this, we seek to address the aforementioned gap by examining scientific research papers that leverage data in their findings. Specifically, we utilize two recent paper-level datasets encompassing data-based papers from the life and social sciences. These datasets are integrated with extensive, large-scale repositories such as the Microsoft Academic Graph (MAG) and PubMed to capture metadata and dynamic features of the authors and research teams involved. To ensure comparability, we apply Coarsened Exact Matching (CEM) techniques (Iacus et al., 2012) and conduct manual verification to pair data-based papers with non-data-based counterparts. This matching enables us to isolate the effects of data utilization on scientific outcomes. Our empirical analysis incorporates a variety of multidimensional metrics, assessing scientific impact, interdisciplinarity (Gates et al., 2019; Stirling, 2007), and disruption (Funk & Owen-Smith, 2017; Park et al., 2023; Wu et al., 2019). By leveraging this rigorous methodology, we aim to directly investigate the influence of data-driven research on the success of scientific endeavors.

This paper contributes to the growing body of literature on the role of data in scientific research, advancing the broader discourse in the science of science and innovation (Fortunato et al., 2018; Wang & Barabási, 2021). Specifically, it provides robust empirical evidence that data-driven papers exhibit higher scientific impact, greater interdisciplinarity, and a stronger potential for disruption within both the life sciences and social sciences. Preliminary comparisons further suggest that data-based papers are often associated with larger, more diverse research teams, characterized by greater academic experience. Additionally, our analysis reveals that while data-based papers are more disruptive, they simultaneously display both higher disruptive and consolidating impacts. This dual effect indicates that data-driven research not only challenges existing paradigms and stimulates the emergence of new areas of inquiry, but also reinforces and expands the current body of knowledge. These findings suggest that data-based research may serve as a bridge between traditional research paradigms and emerging, data-centric methodologies.

The structure of the paper is as follows: Sect. "Background" reviews the relevant literature, providing a comprehensive background to the study. Sect. "Data-based papers" outlines the data sources and the empirical approach. Sect. "Controls and CEM" details the control variables, the preliminary comparisons, and the CEM procedure. Sect. "Innovation measures" describes the measures of scientific impact, interdisciplinarity, and disruption. Sect. "Results" presents the key findings of the analysis, and Sect. "Discussion" discusses the implications, limitations, and potential directions for future research.

Background

Literature review

The rise of data as a central element in scientific research has transformed traditional research processes, fostering novel approaches and enhancing the potential impact of scientific outputs. As more researchers rely on data-intensive methods, studies have increasingly focused on understanding how data availability and accessibility influence citation patterns, innovation, and the broader reach of research (Liu et al., 2023b). In this section, we review recent literature exploring these dynamics, emphasizing the relationship between data sharing and citation impact, the role of data in fostering transparency and reproducibility, and the emergence of data-specific platforms and journals as critical infrastructure for data-driven research.

The correlation between data sharing and citation impact has been a focus of recent research. He and Han (2017) investigated the relationship between the usage of datasets in Dryad and the citation counts of the associated articles. Their study found a strong positive correlation between data usage and article citations, with data-sharing articles consistently outperforming average citation counts across numerous disciplines. This suggests that making research data accessible not only supports transparency and reproducibility but also enhances the visibility and influence of the research. Similarly, Chao (2011) examined the works identified by dataset creators as being closely associated with particular datasets. Du et al. (2021) developed a gold-standard dataset for identifying software mentions in academic papers, using manually annotated PDFs in biomedicine and economics to explore the impact of research software and code as part of the broader research data landscape.

Buneman et al. (2021) further advocated for the routine citation of datasets, arguing that data should be credited in much the same way as traditional publications. They identified two major challenges in data citation: modeling the evolution of credit over time and accurately citing individual parts of datasets rather than treating datasets as monolithic entities. Despite these challenges, they stress the importance of building a standardized data citation framework to support the integration of data into scholarly recognition systems. The rise of data journals represents an additional development in promoting data's role in research. Candela et al. (2015) discuss how data journals enable researchers to publish papers that describe datasets in a way that serves as a citation proxy for the data itself. However, papers that cite data journal articles do not always make direct use of the datasets described, which suggests a gap between data availability and actual data utilization in research.

Several studies have also investigated data-specific platforms to understand how open data infrastructures are used in scholarly work. Robinson-Garcia et al. (2017) analyzed DataCite, a major open data repository, to explore its potential as a bibliometric data source. By examining records from DataCite's OAI API, they offer recommendations for using this platform in bibliometric analysis, particularly for understanding the characteristics and impact of open data on scientific production. Similarly, Qin et al. (2022) analyzed GenBank collaboration networks by integrating metadata from both the data production and publication stages, uncovering new insights into collaboration dynamics and data-sharing practices within the GenBank repository. Koesten et al. (2020) analyzed over a million datasets across repositories such as GitHub to examine how reuse patterns align with features identified in existing literature. Their work contributes to a growing body of knowledge on how datasets are reused across different platforms and

research contexts. Studies have also assessed how datasets relate to the broader academic metrics and societal impact. Acuna et al. (2022) examined tens of thousands of datasets and associated publications on Figshare, linking dataset downloads with authors' h-index, institutional prestige, journal rankings, and citation patterns. Their findings indicate that dataset visibility and accessibility can enhance the academic standing of associated publications and authors.

Yu and Romero (2024) investigated how datasets—particularly those that combine unusual or newly linked data—correlate with scientific and public impact. Using the MICA datasets (Hemphill et al., 2024), they found that unique datasets are often associated with greater visibility and influence in both academic and public spheres. However, a noted limitation of their study was the lack of a control group of non-data-based papers, which limits the ability to draw direct comparisons between data-driven and non-data-driven research.

The adoption of data-driven research is driven by several motivations but also faces significant challenges. Motivations include the potential for increased visibility and citation impact, as data-sharing papers often receive more attention (He & Han, 2017). Funding agencies increasingly require data-sharing plans as a condition for grants, incentivizing researchers to make datasets publicly available (Yin et al., 2021, 2022). Additionally, data-driven approaches facilitate interdisciplinary collaboration, enabling researchers to address complex questions by integrating diverse datasets and methodologies (Gates et al., 2019). For example, repositories like GenBank and ICPSR foster cross-disciplinary projects, enhancing innovation potential. However, researchers face notable challenges in adopting data-driven approaches. Data quality and standardization issues can undermine the reliability of findings, as inconsistent or incomplete datasets may lead to erroneous conclusions. Accessibility barriers, such as restricted access to proprietary datasets or insufficient infrastructure in under-resourced institutions, limit equitable participation in data-driven research. Technical expertise is another hurdle, as advanced data analysis often requires skills in computational tools, machine learning, or statistical modeling, which may not be universally available among researchers. Ethical concerns, including data privacy and the risk of misuse (e.g., in sensitive social science datasets), also pose challenges, requiring robust governance frameworks. These motivations and challenges highlight the complex dynamics influencing the adoption of data-driven research and underscore the need for supportive policies and infrastructure.

Hypothesis

Scientific impact, often measured by citation counts, reflects the extent to which a research article influences subsequent work. Data sharing is widely understood to increase the visibility and credibility of research, thereby amplifying its scientific impact (He & Han, 2017; Liu et al., 2023b). Studies have shown that articles with accessible datasets receive more citations on average than those without, as data-sharing practices allow other researchers to directly engage with, validate, and build upon the original findings (Chao, 2011). Additionally, data-based research articles are increasingly seen as valuable contributions within academic communities, where data-driven methods provide robust, replicable findings that align with the movement towards open science. Given these insights, we hypothesize that data-based papers will demonstrate higher scientific impact compared to non-data-based papers, as the transparency, accessibility, and validation opportunities afforded by data-sharing practices are likely to drive both direct and indirect citations.

H01 Data-based papers exhibit greater scientific impact compared to their control group counterparts.

Interdisciplinarity, which reflects a research paper's capacity to integrate methods, theories, or data from multiple fields, is increasingly recognized as a catalyst for innovation and problem-solving. Data-based research, by its nature, often draws from or contributes to diverse datasets, methodologies, or analytical tools, potentially making it more applicable across disciplinary boundaries (He & Han, 2017). Previous studies suggest that data integration enables research that addresses complex, multifaceted problems which are more likely to appeal to broader research communities, thereby crossing traditional disciplinary lines (Fortunato et al., 2018; Gates et al., 2019). For example, repositories like GenBank or DataCite support diverse scientific collaborations and facilitate the convergence of biology, data science, and computational fields (Robinson-Garcia et al., 2017). We propose that data-based research papers demonstrate greater interdisciplinarity than their control group counterparts, as the shared nature of data encourages collaborations and insights from multiple domains, further expanding the paper's reach and relevance.

H02 Data-based papers exhibit greater interdisciplinarity compared to their control group counterparts.

Disruption in scientific research refers to the extent to which a paper challenges existing paradigms or prompts shifts in understanding within its field (Funk & Owen-Smith, 2017; Park et al., 2023; Wu et al., 2019). Data-driven research may contribute to such paradigm shifts by uncovering new findings, developing novel methods, or revealing connections that may not be apparent through traditional research approaches. Data's potential to disrupt is partly due to its ability to produce reproducible and scalable results, often inspiring other researchers to adopt similar methods or revise their theoretical frameworks in response to new, data-driven insights (Youyou et al., 2023). Disruptive studies often reflect new methodological approaches that are rapidly adopted, spurring further innovation and exploration (Kedrick et al., 2024). For instance, the integration of datasets from multiple disciplines can reveal unanticipated patterns or causal relationships, pushing the boundaries of established research paradigms (Yu & Romero, 2024). Therefore, we hypothesize that data-based papers exhibit higher levels of disruption, as the inclusion of comprehensive data and advanced data-analytic techniques are likely to yield findings that challenge existing scientific conventions.

H03 Data-based papers exhibit greater disruption compared to their control group counterparts.

Data-based papers

The identification and analysis of data-based papers in this study focus on the life sciences and social sciences due to the availability of high-quality, comprehensive datasets that serve as gold standards for data-driven research in these fields. Obtaining robust and well-curated datasets for data-driven research is challenging, as such datasets require extensive metadata, reliable entity extraction, and consistent documentation of data usage. The datasets selected—PubMed Knowledge Graph 2.0 (PKG 2.0) for life sciences and the MICA

dataset from the Inter-university Consortium for Political and Social Research (ICPSR) for social sciences—were chosen for their quality and comprehensive coverage. These datasets enable precise identification of data-based papers and provide rich metadata for analyzing scientific impact, interdisciplinarity, and disruption.

Data-based papers in life science

The identification of data-based papers within the life sciences was conducted using the extensive PubMed dataset, a primary repository of biomedical literature curated by the U.S. National Library of Medicine at the National Institutes of Health. For this study, we employed PubMed Knowledge Graph 2.0 (PKG 2.0), developed by Xu et al. (2024). PKG 2.0 is a sophisticated knowledge graph dataset that integrates over 36 million life science publications from the 1940s to 2024. This dataset provides comprehensive interconnections across various biomedical resources, encompassing biomedical entities, author networks, citation networks, and research projects.

The identification of data-based papers in the life sciences utilized entity extraction through natural language processing methods. Xu et al. (2024) utilized BERN2 (Sung et al., 2022), a robust biomedical text mining tool, to perform both Named Entity Recognition and Named Entity Normalization. BERN2 facilitated the extraction of biomedical entities from titles and abstracts with high precision and efficiency. The authors further implemented a neural network-based NEN approach to achieve enhanced normalization for entities beyond the scope of traditional rule-based methods.

A dictionary-based methodology was applied for dataset knowledge entity extraction, drawing on three primary sources. The first source was the Nucleic Acids Research (NAR) Database (Oxford Academic),[1] followed by the NIH-supported domain-specific repositories provided by the Biomedical Informatics Coordinating Committee (BMIC) as part of the NIH Data Sharing Resources.[2] This repository encompasses both domain-specific and general data-sharing repositories, each annotated with specific properties. The third source was the European Bioinformatics Institute (EMBL-EBI) database,[3] which is known for maintaining an extensive array of freely accessible molecular data resources. Each of these sources was retrieved in its latest version, collectively identifying a total of 1,809 unique datasets, leading to the identification of 469,551 relevant papers. To enhance accuracy, we excluded the PubMed dataset itself from consideration, ultimately yielding a final dataset comprising 247,847 papers and 1,808 unique data sources within the life sciences.

Data-based papers in social science

In social science field, we employed a dataset developed as part of the MICA project (Measuring the Impact of Curation Actions), led by researchers at the Inter-university Consortium for Political and Social Research (ICPSR) and documented by Hemphill et al. (2024). The MICA dataset was specifically designed to assess the impact of archival and data curation decisions on data reuse. It was constructed from study-level metadata, citation records, and curation records available through ICPSR, a major social science data

[1] https://www.oxfordjournals.org/nar/database/a/.
[2] https://www.nlm.nih.gov/NIHbmic/bmic-about.html.
[3] https://www.ebi.ac.uk/services.

archive hosted by the University of Michigan. The ICPSR bibliography of data-related literature, a growing repository of literature utilizing ICPSR data, was foundational to this dataset. Initially supported by the National Science Foundation (Award 9,977,984), the bibliography has since been funded for over two decades by ICPSR member contributions and various federally and foundation-supported topical archives housed within ICPSR.

To ensure comprehensive inclusion, Hemphill et al. (2024) conducted a detailed search across the Dimensions AI database for all ICPSR study names, unique ICPSR identifiers, and Digital Object Identifiers using a full-text index. The resulting dataset provides a robust description of 10,605 social science research datasets and the corresponding usage contexts within 94,755 scholarly publications, spanning from 1963 to 2022.

To obtain enriched citation and metadata information, these data-based publications were mapped to entries in the Microsoft Academic Graph (MAG) database using Digital Object Identifiers and publication titles. The Microsoft Academic Graph (Wang et al., 2020) is recognized as one of the most extensive bibliometric resources, encompassing over 200 million scholarly records—including journal articles, conference proceedings, and preprints—dating from 1800 to 2021. For this analysis, we utilized the MAG release dated December 6, 2021. Through this mapping, we identified a subset of 57,913 publications that are data-driven and specific to social science research.

Figure 1 presents the distribution of datasets and data-centric publications within our sample. In the life sciences, the five most frequently utilized databases include the Gene Ontology, KEGG, GenBank, Gene Expression Omnibus, and Protein Data Bank, each cited in over 1000 instances within our dataset. Similarly, in the social sciences, the American National Election Study, Uniform Crime Reporting Program Data, Monitoring the Future, Current Population Survey, and National Health and Nutrition Examination Survey represent the five most frequently cited datasets, each also appearing in over 1,000 publications in our sample. Additionally, the trend observed suggests a steady increase in the number of data-centric publications within the life sciences over time.

Controls and CEM

Control variables

To account for potential differences between data-based papers and other types of publications in both life sciences and social sciences, we incorporate a set of control variables in our analysis. These variables are essential, as they may significantly impact measures of scientific innovation; thus, careful assessment and control of these variables are warranted.

First, we consider the publication year of each paper. Earlier publications may exhibit higher disruptiveness, lower interdisciplinarity, and fewer citations within comparable time windows, as innovation patterns in science evolve over time. For instance, Park et al. (2023) found that scientific papers have become less disruptive over time, potentially due to the consolidation of established paradigms in mature fields. Conversely, Yang (2025) observed that interdisciplinarity has increased in recent decades, reflecting the growing integration of diverse methodologies and data-driven approaches in modern research. Additionally, when measured within fixed time windows, citations have increased for more recent papers, likely due to the expansion of scientific literature and improved accessibility through digital platforms (Fortunato et al., 2018).

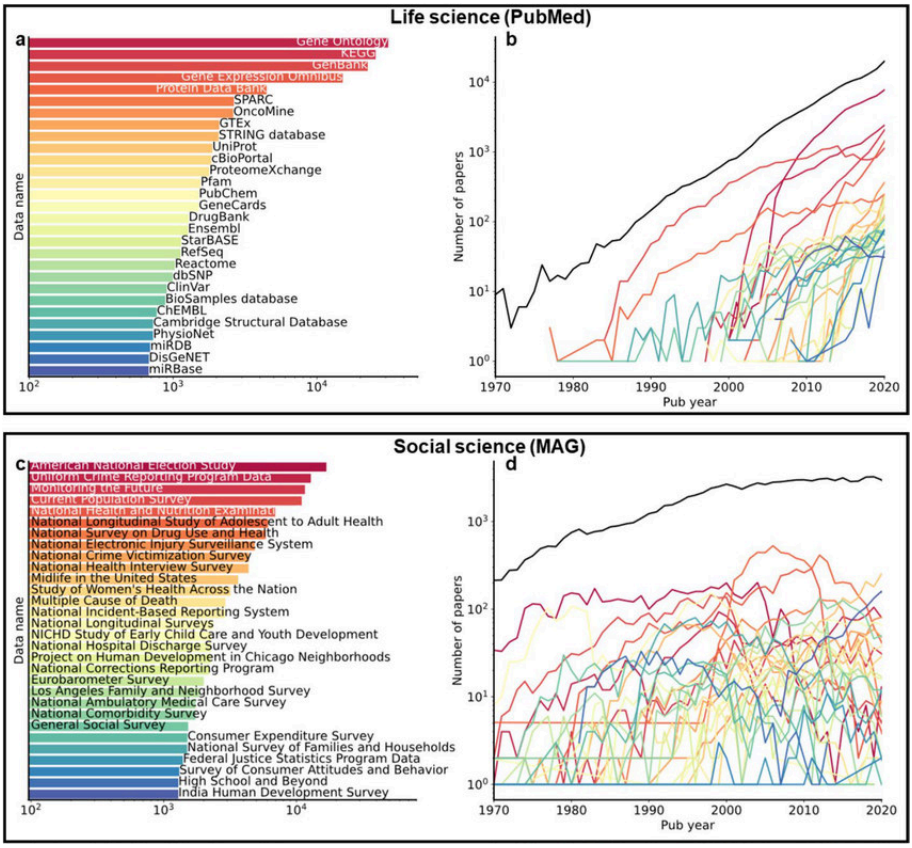


Fig. 1 Trends and distributions of datasets and data-based publications in our sample. **a** The top 30 most frequently used datasets in life sciences. **b** The number of data-based publications in life sciences, with the black line representing the overall number of data-based papers and the other lines showing the number of data-based papers corresponding to the datasets in (**a**). **c** The top 30 most frequently used datasets in social sciences. **d** The number of data-based publications in social science, with the black line indicating the overall number of data-based papers and the other lines representing the number of data-based papers corresponding to the datasets in (**c**)

We also account for the number of authors per paper as a measure of team size. Smaller teams are often associated with greater disruptive potential but lower citation impact, as demonstrated by Wu et al. (2019). Their study found that smaller teams tend to produce highly disruptive work by focusing on high-risk ideas that challenge existing paradigms, often with fewer resources or collaborative constraints. However, these papers may receive fewer citations due to their specialized nature, which may limit their immediate appeal to broader audiences compared to the incremental, widely cited work often produced by larger, more resource-rich teams.

Beyond team size, we integrate several metrics related to team diversity, utilizing dynamic career data from MAG and PubMed. Both databases provide disambiguated author data, enabling us to track authors' career trajectories and other relevant information. We also employ gender data from SciSciNet (Lin et al., 2023) and PubMed to construct gender-based team variables. Prior research suggests that gender-diverse teams are

more likely to generate novel ideas (Yang et al., 2022). We, therefore, introduce a variable termed "gender entropy," calculated as follows:

$$Genderentropy = -p_f ln\left(p_f\right) - (1 - p_f) ln\left(1 - p_f\right)$$

where p_f represents the probability of female authors in team. This metric ranges from 0 to 1, with higher values signifying greater gender diversity. Additionally, we calculate the proportion of female authors for each team.

To further control for team composition, we examine the academic status of each author as represented by their h-index. (Yang et al., 2025). Research suggests that flatter teams are more likely to produce disruptive science, while more hierarchical teams often contribute to cumulative scientific development (Xu et al., 2022). We calculate a "team hierarchy" measure using the Gini index of h-index values across team members, as defined by:

$$Teamhierarchy = \frac{\sum_{m=1}^{N} \sum_{n=1}^{N} |h_m - h_n|}{2N^2 \overline{h}}$$

where h_m represents the dynamic h index of member m, based on their publications and citations up to the year prior to the focal paper's publication. and $\overline{h}$ denotes the average dynamic h index of all team members. This Gini index ranges from 0 to 1, with higher values indicating more hierarchical teams.

Additionally, we calculate the average h-index and career age of team members to reflect the team's collective academic experience. Career age is defined as the number of years since an author's first publication in the PubMed (life science) or MAG (social science) dataset.

Since citation-based measures heavily depend on the references included in a paper, we control for the number of references as a proxy for the scope and depth of the literature reviewed (Leydesdorff et al., 2019). Additionally, we account for journal tier, as a proxy for publication prestige. For PubMed data, journal tier is derived from SciMago rankings, while in MAG, it is based on impact factor. Journal tiers are categorized into four quantiles, and the highest annual rank of each journal is used.

Finally, while broad disciplinary fields are already controlled for, we also account for subfield variations within both the life and social sciences. In the life sciences, we use the Medical Subject Headings (MeSH) tree structure to identify subfields. For the social sciences, we employ the second-level MAG classification to delineate subfield distinctions.

Preliminary comparison

We conduct a comparative analysis of data-based publications versus other types of publications, with a focus on a series of control variables. Here, "other papers" denotes all publications not identified as data-based within the same field; these are not considered control groups per se. Nonetheless, this comparison yields preliminary insights into the distinct characteristics and patterns associated with data-based publications.

Figure 2 presents the comparison between data-based and non-data-based papers in the life sciences (PubMed) and social sciences (MAG) along key dimensions: publication year, team size, gender diversity, proportion of female team members, team hierarchy, average h-index of team members, average career age, number of references, and journal impact quantiles. In analyzing social science publications, we use first-level field labels from

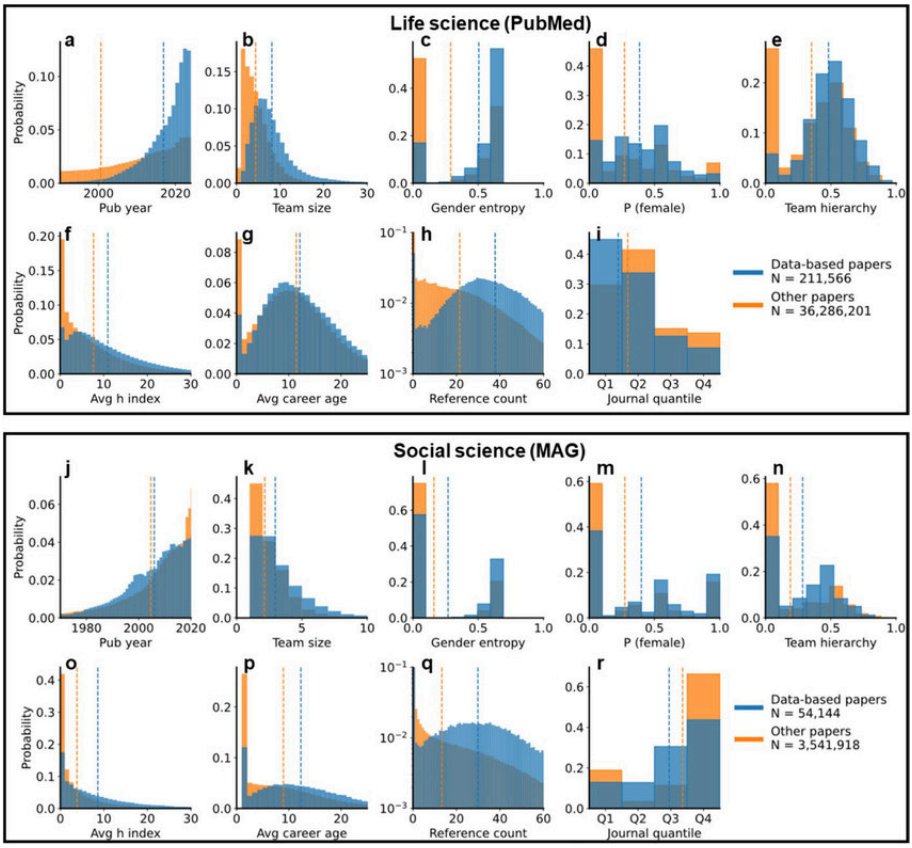


Fig. 2 Comparison of data-based versus other publications in life and social sciences. **a**–**i** Analysis of data-based papers versus other life science papers in PubMed, comparing: **a** publication year, **b** team size, **c** gender diversity (gender entropy), **d** proportion of female team members, **e** team hierarchy, **f** average h-index of team members, **g** average career age, **h** number of references, and **i** journal impact quantiles. **j**–**r** Analysis of data-based papers versus other social science papers in MAG, comparing: **j** publication year, **k** team size, **l** gender diversity (gender entropy), **m** proportion of female team members, **n** team hierarchy, **o** average h-index, **p** average career age, **q** number of references, and **r** journal impact quantiles. Vertical lines indicate the average values of data-based papers and other papers

MAG to define the scope of papers while excluding others. In PubMed, however, we do not apply this limitation, as the majority of indexed publications pertain to the life sciences.

In the life sciences, data-based publications reveal several notable distinctions compared to other papers. On average, data-based papers have more recent publication dates (mean publication year of 2017 vs. 2001; T-test, $p < 0.001$), larger collaborative teams (mean team size of 8.1 vs. 4.4; T-test, $p < 0.001$), and higher levels of gender diversity within teams (mean gender diversity of 0.50 vs. 0.29; T-test, $p < 0.001$). Additionally, data-based publications show a greater proportion of female contributors (mean proportion of 0.39 vs. 0.27; T-test, $p < 0.001$), higher hierarchical structuring within teams (mean team hierarchy of 0.48 vs. 0.35; T-test, $p < 0.001$), and a higher average h-index among team members (mean h-index of 11 vs. 8; T-test, $p < 0.001$). These papers are also produced by teams with greater average career experience (mean career age of 12 years vs. 11 years; T-test,

$p < 0.001$) and include a higher number of cited references (mean of 37.8 references vs. 21.5; T-test, $p < 0.001$).

In the social sciences, data-based publications similarly exhibit distinct characteristics when compared to other papers. On average, data-based papers are more recent (mean publication year of 2006 vs. 2005; T-test, $p < 0.001$), have larger collaborative teams (mean team size of 3.0 vs. 2.2; T-test, $p < 0.001$), and show greater gender diversity within teams (mean gender diversity of 0.27 vs. 0.16; T-test, $p < 0.001$). These publications also demonstrate a higher proportion of female contributors (mean proportion of 0.40 vs. 0.27; T-test, $p < 0.001$), increased hierarchical structuring within teams (mean team hierarchy of 0.28 vs. 0.19; T-test, $p < 0.001$), and a higher average h-index among team members (mean h-index of 9 vs. 4; T-test, $p < 0.001$). Furthermore, teams producing data-based papers have greater average career experience (mean career age of 12 years vs. 9 years; T-test, $p < 0.001$) and cite a higher number of references (mean of 30.0 references vs. 13.2; T-test, $p < 0.001$).

Overall, these findings indicate that data-based publications are typically associated with larger, more diverse teams with greater academic experience. These factors, therefore, warrant careful control in subsequent analyses to accurately assess the innovative potential of data-based publications.

Coarsened exact matching

To ensure comparability between data-based publications and control group papers prior to conducting regression analysis, we employed Coarsened Exact Matching (CEM) (Iacus et al., 2012). This advanced matching technique enhances the initial similarity between data-based and non-data-based papers. The CEM algorithm follows three principal steps: (1) Temporarily coarsening each control variable X to facilitate matching; (2) Sorting all observations into strata that share identical value of the coarsened X; and (3) Pruning units from the dataset that do not contain at least one treated and one control unit within any stratum. We implemented a one-to-one matching procedure, pairing each data-based publication with a control group counterpart to ensure numerical balance within the sample.

A comprehensive set of features was used to control for publication year and subfield, team size and diversity, author experience and academic status, demographic characteristics, and knowledge foundations. These variables included team size, gender diversity (gender entropy), proportion of female team members, team hierarchy, average h-index of team members, average career age, number of references, and journal impact quantiles. All control variables were selected ex-ante, meaning they could be determined at the time of publication; future impact metrics were not included in the matching process. Publication year and subfield used exact matching, while other variables were coarsened into 10 to 20 coarsened strata to facilitate matching precision.

To further ensure rigor, we conducted a manual verification of control group papers by examining their titles and abstracts to confirm that no unidentified data-based papers were inadvertently included. In cases where a control paper was found to be data-based, we reran the CEM for that specific treatment paper, ensuring that the control group exclusively comprised non-data-based publications and thereby improving the accuracy of our subsequent analyses.

To ensure the rigor of the control group classification and avoid misclassification of data-based papers as non-data-based, we implemented a systematic manual verification process for the control group papers. This process was designed to confirm that control papers did not rely on publicly accessible datasets, internal clinical data, or other data

sources that could qualify them as data-based under our criteria. The verification procedure involved the following steps:

(1) Two independent researchers reviewed the titles and abstracts of all control group papers identified through CEM. The review focused on identifying explicit mentions of dataset usage (e.g., references to repositories like GenBank, ICPSR, or EMBL-EBI) or descriptions of data-driven methodologies (e.g., large-scale genomic analyses, statistical modeling of public datasets). Papers with ambiguous language (e.g., mentioning "data" without specifying a source) were flagged for further inspection.
(2) For papers flagged as ambiguous during the title and abstract review, we accessed the full text (when available) to examine the methods section for evidence of data usage. This step was particularly critical in the life sciences, where retrospective studies might use small-scale internal clinical data (e.g., hospital case records) or publicly accessible datasets not explicitly mentioned in abstracts. We considered a paper data-based if it used any systematically collected dataset, whether public or internal, that was analyzed statistically or computationally to derive results.
(3) To avoid misclassification, we established clear criteria for identifying data-based papers. In the life sciences, experimental studies generating original data (e.g., lab-based sequencing results, cell-based measurements) were classified as data-based only if the data were deposited in a publicly accessible repository (e.g., Gene Expression Omnibus, Protein Data Bank) or if the study explicitly described the use of such repositories. Studies relying solely on small-scale, non-deposited, or ad-hoc experimental data (e.g., lab-specific measurements not shared publicly) were classified as non-data-based to focus on the impact of shared or reusable data. In the social sciences, small-scale survey studies based on non-public questionnaires or interviews with limited sample sizes (e.g., fewer than 100 participants) were classified as non-data-based unless they utilized publicly accessible datasets or were part of a larger, curated repository like ICPSR. This distinction ensured that our definition of "data-based" emphasized research leveraging shared or large-scale data resources.
(4) If a control paper was identified as data-based during the verification process, it was excluded, and the CEM algorithm was rerun for the corresponding treatment paper to identify a new control match. This iterative process continued until all control papers were confirmed as non-data-based. To assess the reliability of the manual verification, a third researcher independently reviewed a random 10% sample of control papers, achieving an inter-rater agreement of 92% (Cohen's kappa = 0.88), indicating high consistency.
(5) In the life sciences, we paid particular attention to retrospective studies that might use internal clinical data (e.g., hospital records) not explicitly disclosed in abstracts. For such cases, we cross-referenced the paper's institutional affiliations with known clinical data repositories and checked for mentions of data-sharing compliance (e.g., NIH data-sharing policies). If no evidence of dataset usage was found, the paper was retained as non-data-based.

In the life sciences, we successfully matched 189,557 data-based publications with an equal number of control group publications. In the social sciences, we achieved successful matching for 45,244 data-based publications and their control group counterparts. An examination of the matched groups' characteristics revealed that, prior to

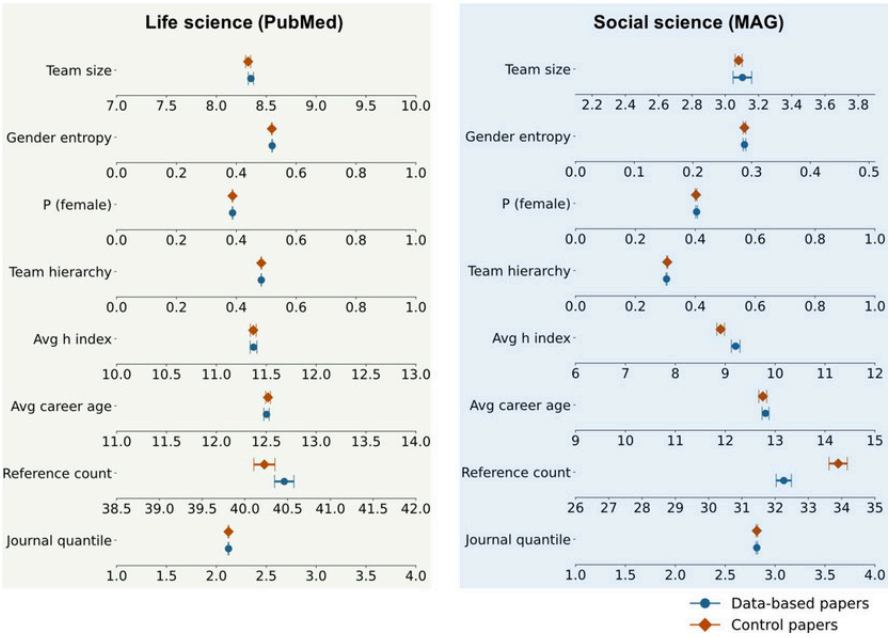


Fig. 3 Post-CEM feature comparisons between data-based papers and control group. Eight characteristics were compared, controlling for subfield differences. The features analyzed (from top to bottom) include: team size, gender diversity (gender entropy), proportion of female team members, team hierarchy, average h-index of team members, average career age, number of references, and journal impact quantiles. No significant differences were observed between the two groups across any measured dimension following CEM. Error bars represent 95% confidence intervals

publication, data-based and non-data-based papers were statistically indistinguishable across all measured dimensions (Fig. 3).

Innovation measures

Interdisciplinarity

Following Stirling (2007) framework, this investigation dissects interdisciplinarity into three primary dimensions: variety, entropy, and disparity. Additionally, we employ the Rao-Stirling index to measure overall interdisciplinarity. The calculation of interdisciplinarity values in MAG dataset relies on 292 defined secondary subfields. For PubMed, which lacks an equivalent subfield classification for individual papers, we utilize the Medical Subject Headings (MeSH) hierarchical structure[4] to ascertain paper disciplines. MeSH descriptors are categorized into 16 broad groups, such as anatomic terms (Category A), organisms (Category B), and diseases (Category C). Each category is organized into subcategories with descriptors that range hierarchically from general to specific, extending up to thirteen levels of specificity. Due

[4] https://meshb.nlm.nih.gov/treeView.

to this branched structure, these categories are often termed "trees." Specifically, we use the second level of the MeSH hierarchy (e.g., "Inorganic Chemicals D01"). For each focal paper, we focus on the fields of its references, and let each reference papers' field duplicate, thereby forming a merged field distribution (Fig. 4).

We quantify interdisciplinarity by analyzing the distribution of disciplines among the references cited in scholarly articles. For a given paper with a set of references $R = \{r1, r2, \dots\}$, the fields of the references are mapped into a field distribution vector $F = \{field_{r1}, field_{r2}, \dots, field_{rk}, \dots\}$, where $field_{rk}$ reflects the field of reference r_k.

Interdisciplinarity Variety is represented by the number of distinct fields cited in the references of a focal paper and is formally defined as:

$$Variety = ln(\sum_{r \in R} field_r)$$

Interdisciplinarity Entropy captures the diversity of disciplines among the references of a focal paper and is calculated using Shannon entropy:

$$Entropy = -\sum p_m ln(p_m)$$

where $m \in F$ represents a field of the references of a focal paper, p_m denotes the probability of a reference belonging to field m.

Interdisciplinarity Disparity reflects the maximum conceptual distance between disciplines referenced by a focal paper and is defined as:

$$Disparity = Max\{d_{mn} | m, n \in F_i\}$$

where d_{mn} represents the distance between fields m and n, calculated based on the entire set of papers and citation networks in the MAG or the PubMed dataset. Specifically,

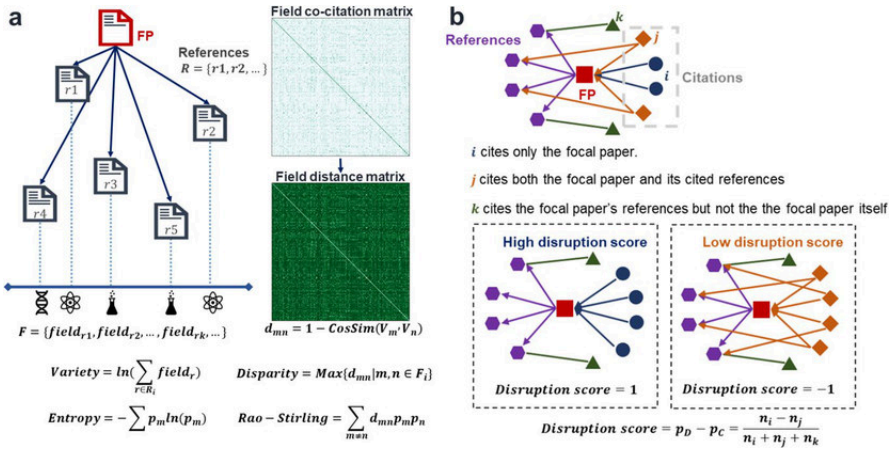


Fig. 4 Measuring interdisciplinarity and disruption in scientific papers. **a** Interdisciplinarity of papers is assessed using metrics including variety, entropy, disparity, and the Rao-Stirling index, based on the disciplinary fields of referenced papers, identified by second-level subfields in MAG or second-level MeSH categories in PubMed. **b** The CD index is calculated to quantify disruption, representing the relative proportion of disruptive citations (citations to the focal paper without its references). Additionally, disruptive citations n_i and consolidating citations n_j are also analyzed as dependent variables

considering all papers in field m denoted by$M = \{i|field_i \in m\}$, a large-scale disciplinary co-citation matrix is constructed based on citation patterns. This matrix forms a co-citation distribution vector V_m between field m and the other fields is considered, reflecting the distribution of knowledge interactions between different disciplines. The cosine distance is calculated to determine the distance d_{mn} between fields m and n is computed as:$d_{mn} = 1 - CosSim(V_m, V_n)$.

Rao-Stirling Index is a comprehensive measure of interdisciplinarity, defined as:

$$Rao - Stirling = \sum_{m \neq n} d_{mnt} p_m p_n$$

where p_m denotes the fraction of references in discipline m, and d_{mnt} is the distance between field m and n in year t.

Disruptiveness

This study employs the CD index (Funk & Owen-Smith, 2017; Park et al., 2023; Wu et al., 2019), a quantitative measure developed to assess the degree of disruption in scientific work. The CD index is formally defined as:

$$CDindex = p_D - p_C = \frac{n_i - n_j}{n_i + n_j + n_k}$$

where n_i represents the number of future papers citing the focal paper (FP) that do not reference its cited sources, n_j denotes the number of future papers citing FP and its references, and n_k represents the number of papers that cite the FP's references without citing the FP itself. A lower CD index suggests alignment with established knowledge, while a higher CD index signifies a paper with the potential to drive paradigm shifts.

A lower CD index value indicates that the focal paper cites references that are also widely cited by subsequent papers referencing the FP. This trend implies the focal paper's role in consolidating existing paradigms, aligning with and reinforcing established knowledge. Conversely, a high CD index is indicative of a disruptive contribution, suggesting that the paper may catalyze new fields or redefine existing knowledge boundaries (Yang et al., 2024). Thus, papers with higher CD index values are classified as highly disruptive, reflecting a potential to instigate paradigm shifts, whereas lower values suggest consolidation within existing research paradigms.

By applying the CD index, this study aims to elucidate the disruptive impact of scientific research, shedding light on knowledge dissemination patterns and paradigm shifts within scientific domains. Nevertheless, the CD index has limitations; it is heavily influenced by the number of references in a focal paper (Yang et al., 2024). This potential bias is mitigated in this study by employing CEM to pair treatment papers with control papers that have similar reference counts.

In addition to the CD index, which primarily reflects the relative proportion of disruptive citations (citations to the FP without its references), this analysis also examines disruptive citations n_i and consolidating citations n_j as dependent variables. These variables capture the disruptive and consolidative impacts of focal papers, respectively. Should the treatment effect increase the disruptive impact, it would suggest that data-based papers exert a more disruptive influence on their fields.

Results

Data-based papers have higher scientific impact

We employed multivariable Ordinary Least Squares (OLS) regression to assess whether data-based publications exhibit greater scientific impact compared to their control group counterparts. Figure 5 and Appendix Table 1 presents the results, where the sample comprises both treatment (data-based) papers and matched control papers from the CEM procedure. Control variables are included in the regression model, along with fixed effects for publication year, subfield, and journal impact quantile, to account for temporal and disciplinary heterogeneity. Analyses were conducted separately for papers in the life sciences and social sciences.

The findings consistently indicate that data-based papers have a higher scientific impact than control group papers. In the life sciences, data-based publications exhibit a 6.2% increase in long-term citation counts and a 6.4% increase in 5-year citation counts relative to the control group. In the social sciences, data-based papers show an even more pronounced impact, with 36.1% higher long-term citation counts and 20.1% higher 5-year citation counts compared to control publications. Additionally, we examined the media attention data-based papers receive, specifically in terms of News mentions and Tweets (X), based on data from SciSciNet. For social science papers, data-based publications receive 1.6% more News mentions and 10.1% more mentions on Twitter than non-data-based papers. These results highlight the enhanced visibility and impact of data-based

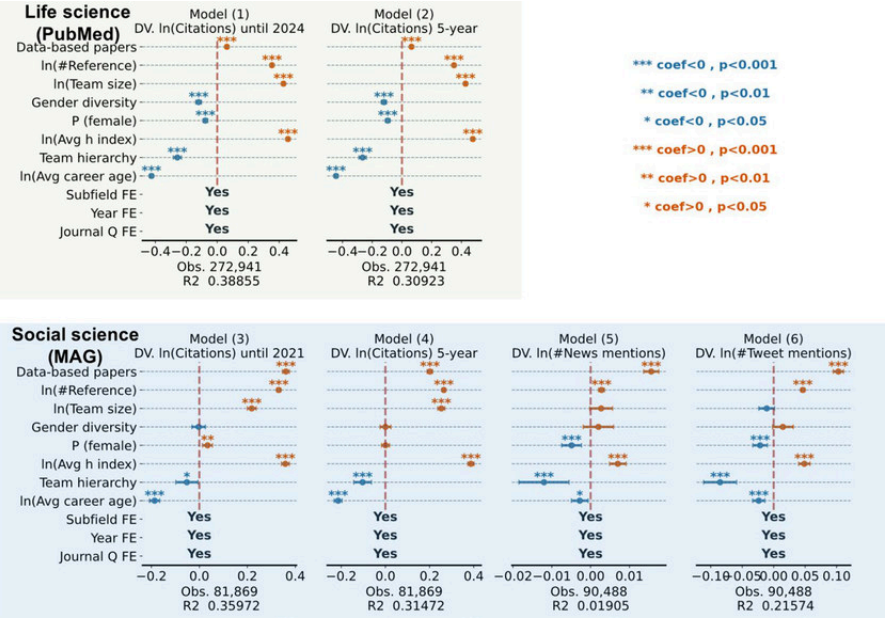


Fig. 5 Effect of the data-based papers on the scientific impact of papers. Models (1–6) utilize Ordinary Least Squares (OLS) regression models. In all models, the coefficients of Data-based papers demonstrate significant positive effects. Error bars depict 95% CI based on robust standard errors. *p < 0.05; **p-value < 0.01; ***p-value < 0.001

research, emphasizing the value of such papers in contributing to scientific knowledge and generating broader societal influence.

Data-based papers are more interdisciplinary

We then conducted multivariable OLS regression to examine whether data-based publications demonstrate greater interdisciplinarity compared to their control group counterparts. Figure 6 (Models 1–4) and Appendix Table 2 presents the results for life sciences, where interdisciplinarity measures were derived using the MeSH tree structure of the reference papers. The findings indicate that data-based papers exhibit significantly higher interdisciplinary values across all measured dimensions—variety, entropy, disparity, and Rao-Stirling—compared to non-data-based papers. Figure 6 (Models 5–8) and Appendix Table 3 displays the results for the social sciences, where interdisciplinarity was assessed using the second-level MAG subfield classification of reference papers. Similar to the life sciences, data-based papers in the social sciences also demonstrate significantly higher levels of interdisciplinarity across all metrics, with effect sizes notably larger than those observed in the life sciences.

These results suggest that data-based research is more interdisciplinary than non-data-based research, both in the life sciences and social sciences. The higher levels of interdisciplinarity observed in data-based publications may reflect the nature of data-driven research, which often requires the integration of diverse methodologies, perspectives, and theoretical frameworks from multiple fields. This is particularly true for complex research

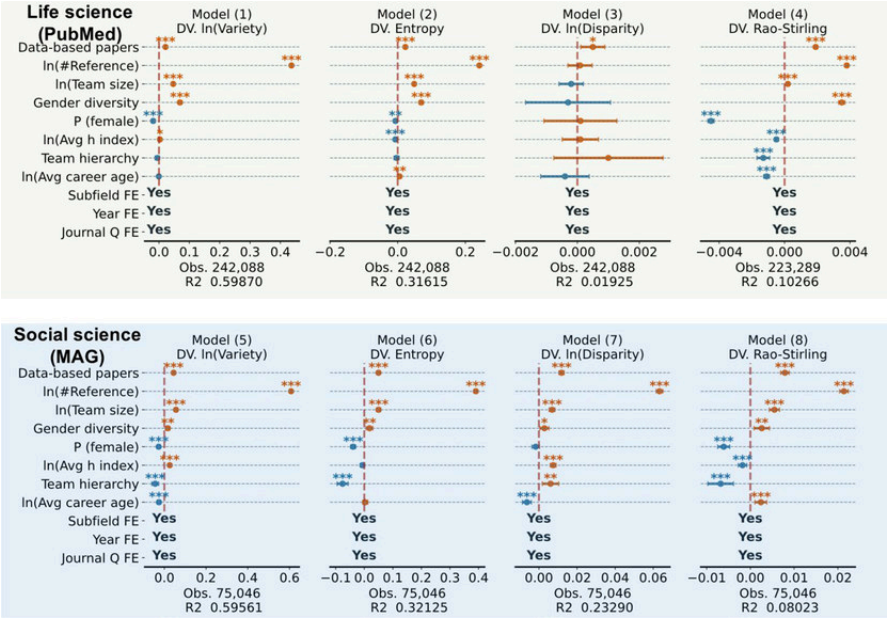


Fig. 6 Effect of the data-based papers on the interdisciplinarity of papers. Models (1–8) utilize Ordinary Least Squares (OLS) regression models. In all models, the coefficients of Data-based papers demonstrate significant positive effects. Error bars depict 95% CI based on robust standard errors. *p < 0.05; **p-value < 0.01; ***p-value < 0.001

questions that necessitate the use of large datasets, advanced analytical techniques, and collaboration across disciplinary boundaries.

Data-based papers are more disruptive

We then performed multivariable OLS regression analyses to investigate whether data-based publications exhibit greater disruptive potential compared to their control group counterparts. The results for the life sciences are presented in Fig. 7 and Appendix Tables 4, 5, Models (1–3), which reveal that data-based papers have significantly higher CD index values than control papers. The modest effect size observed is likely due to the CD index being centered around zero. To further explore this, we analyzed the impact of data-based papers on two key outcomes: the number of forward disruptive citations (n_i) and the number of forward consolidating citations (n_j). Disruptive citations (n_i) represent future papers citing the focal paper without referencing its cited sources, while consolidating citations (n_j) denote papers that cite both the focal paper and its references. Our findings indicate that data-based papers receive both higher disruptive and higher consolidating citations compared to control papers. Furthermore, the effect size for disruptive citations (0.08) is notably larger than that for consolidating citations (0.02), with both effects being statistically significant at the 0.001 level.

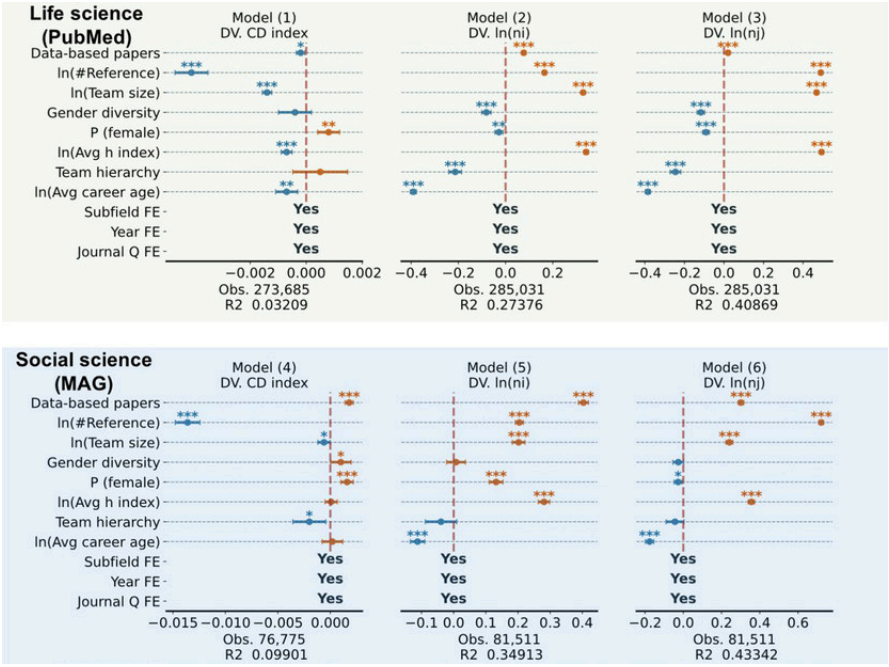


Fig. 7 Effect of the data-based papers on the disruptive potentials of papers. Models (1–6) utilize Ordinary Least Squares (OLS) regression models. In all models, the coefficients of Data-based papers demonstrate significant positive effects. Error bars depict 95% CI based on robust standard errors. *p < 0.05; **p-value < 0.01; ***p-value < 0.001

Models (4–6) present the results for the social sciences, which also show that data-based papers exhibit significantly higher CD index values than their control counterparts. Similarly, data-based papers in the social sciences receive both higher disruptive and higher consolidating citations than control papers. Again, the effect size for disruptive citations (0.40) is larger than that for consolidating citations (0.30), with both differences being statistically significant at the 0.001 level.

Taken together, these findings indicate that data-based papers have a more substantial disruptive and consolidating impact than control papers. The larger effect size for disruptive citations suggests that data-based papers have a higher potential for inducing paradigm shifts, emphasizing their greater disruptive potential relative to traditional publications. This suggests that data-based research may play a more significant role in challenging existing paradigms and fostering the emergence of new areas of inquiry. The observed higher consolidating citations for data-based papers also indicate that, while these papers are disruptive, they simultaneously reinforce and expand the existing body of knowledge, suggesting that they may bridge gaps between traditional research and novel, data-driven approaches.

Discussion

This paper contributes to the expanding literature on the science of science by providing robust empirical evidence on the transformative role of data-driven research in advancing scientific progress. Our findings demonstrate that data-based papers in the life and social sciences exhibit significantly higher scientific impact, interdisciplinarity, and disruption compared to their non-data-based counterparts, aligning with and extending recent trends in the research status quo. The increasing prevalence of data-sharing platforms, such as PubMed, ICPSR, and DataCite, and the growing emphasis on open science practices (Hemphill et al., 2024) underscore the timeliness of our results. These findings resonate with current efforts to enhance research transparency and reproducibility, as evidenced by studies showing that data-sharing correlates with higher citation counts (He & Han, 2017) and fosters cross-disciplinary collaboration. By quantifying the dual role of data-driven research in both challenging existing paradigms and reinforcing knowledge bases, our work highlights data's pivotal role in shaping the evolution of scientific fields, bridging traditional research paradigms with emerging, data-centric methodologies.

From a practical standpoint, our results advocate for continued investment in data-sharing infrastructure and policies, which are increasingly central to modern scientific ecosystems. The higher interdisciplinarity of data-based papers suggests that funding agencies should prioritize support for cross-disciplinary collaborations that leverage large, shared datasets to address complex, multifaceted research questions. For instance, repositories like GenBank and the Gene Expression Omnibus have enabled breakthroughs in genomics and bioinformatics by facilitating data reuse across disciplines (Qin et al., 2022). Similarly, in the social sciences, ICPSR's curated datasets have supported impactful research on societal issues, as seen in the frequent use of datasets like the American National Election Study (Hemphill et al., 2024). Researchers should leverage these resources to enhance the impact and visibility of their work, while journals could incorporate metrics of interdisciplinarity and disruptive potential into their evaluation criteria, particularly for data-intensive fields. Institutions, in turn, should foster data-centric research cultures to boost global recognition

and competitiveness, aligning with the growing adoption of data-driven methodologies in leading research environments (Robinson-Garcia et al., 2017).

Despite offering valuable insights, this study has several limitations. First, the analysis relies on two large datasets from the life and social sciences, which may not fully represent other academic disciplines, such as the humanities or emerging interdisciplinary fields. This limitation restricts the generalizability of the findings to all areas of scientific inquiry. Second, while Coarsened Exact Matching (CEM) effectively controls for observable characteristics when pairing data-based papers with non-data-based counterparts, it cannot account for unobserved factors, such as specific research methodologies or content-level features, that may also influence scientific outcomes. Furthermore, the measurement of disruption through the CD index, which is based on citation patterns, may not fully capture the broader impacts of paradigm shifts that occur outside citation metrics, such as the creation of new research networks or methodologies (Yang et al., 2024, 2025a, 2025b). Additionally, although the datasets used in this study are among the most recent (Hemphill et al., 2024; Xu et al., 2024), they do not capture all data-driven research papers. There remains a need for collaborative efforts to build larger, more comprehensive datasets on data-driven research and better understand how datasets influence scientific innovation. While manual verification in the CEM process focused on titles and abstracts, the potential for bias in the control group remains, as additional factors may not have been considered.

Looking ahead, several promising avenues for future research should be explored. Expanding the scope of data sources to include other disciplines, such as engineering or the humanities, would offer a more comprehensive understanding of how data-driven research influences innovation across a broader spectrum of scientific fields. Further studies could also examine the long-term impact of data-driven research by tracking its influence on scientific progress over extended time periods. Enhancing the measurement of disruption, potentially by incorporating alternative metrics such as the formation of new research fields or collaborations, would provide a deeper understanding of its broader impact. Additionally, exploring the role of research team composition—such as team size, diversity, and interdisciplinarity—could illuminate how collaborative practices contribute to the success of data-driven research. Finally, investigating the relationship between data-driven research and public policy, including funding and data-sharing initiatives, could provide valuable insights into how institutional factors shape scientific progress.

Appendix A: Tables

See Tables 1, 2, 3, 4, 5.

Table 1 Effect of the data-based papers on the scientific impact of papers (OLS regression)

Models	(1)	(2)	(3)	(4)	(5)	(6)
Field/data	Life science (PubMed)		Social science (MAG)			
D.V	ln(Citations) Until 2024	ln(Citations) 5-year	ln(Citations) Until 2021	ln(Citations) 5-year	ln(#News mentions)	ln(#Tweet mentions)
Data-based papers	0.0623*** (0.0036)	0.0642*** (0.0035)	0.3608*** (0.0075)	0.2014*** (0.0065)	0.0156*** (0.0010)	0.1028*** (0.0042)
ln(#Ref)	0.3537*** (0.0031)	0.3498*** (0.0030)	0.3311*** (0.0044)	0.2644*** (0.0037)	0.0028*** (0.0004)	0.0463*** (0.0017)
ln(Team size)	0.4278*** (0.0040)	0.4277*** (0.0039)	0.2185*** (0.0096)	0.2530*** (0.0085)	0.0027. (0.0015)	− 0.0110. (0.0064)
Gender diversity	− 0.1201*** (0.0096)	− 0.1212*** (0.0092)	− 0.0027 (0.0143)	0.0004 (0.0126)	0.0020 (0.0020)	0.0146. (0.0085)
P (female)	− 0.0763*** (0.0083)	− 0.0955*** (0.0080)	0.0343** (0.0106)	0.0005 (0.0092)	− 0.0049*** (0.0013)	− 0.0216*** (0.0059)
ln(Avg h index)	0.4569*** (0.0048)	0.4763*** (0.0046)	0.3586*** (0.0086)	0.3871*** (0.0074)	0.0070*** (0.0011)	0.0487*** (0.0046)
Team hierarchy	− 0.2575*** (0.0135)	− 0.2647*** (0.0131)	− 0.0518* (0.0236)	− 0.1032*** (0.0205)	− 0.0120*** (0.0033)	− 0.0851*** (0.0134)
ln(Avg career age)	− 0.4247*** (0.0065)	− 0.4436*** (0.0062)	− 0.1863*** (0.0109)	− 0.2138*** (0.0092)	− 0.0028* (0.0011)	− 0.0237*** (0.0050)
Subfield FE	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Journal Q FE	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.38855	0.30923	0.35972	0.31472	0.01905	0.21574

Robust standard errors are reported in parentheses. *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$

Table 2 Effect of the data-based papers on the interdisciplinarity of papers (OLS regression)

Models	(1)	(2)	(3)	(4)
Field/Data	Life science (PubMed)			
D.V	ln(Variety)	Entropy	ln(Disparity)	Rao-Stirling
Data-based papers	0.0214*** (0.0009)	0.0225*** (0.0009)	0.0005* (0.0002)	0.0019*** (4.73e−5)
ln(#Ref)	0.4366*** (0.0011)	0.2415*** (0.0012)	8.61e−5 (0.0002)	0.0038*** (4.66e−5)
ln(Team size)	0.0468*** (0.0009)	0.0488*** (0.0009)	− 0.0002 (0.0002)	0.0002*** (4.62e−5)
Gender diversity	0.0691*** (0.0027)	0.0695*** (0.0028)	− 0.0003 (0.0007)	0.0035*** (0.0001)
P (female)	− 0.0192*** (0.0024)	− 0.0067** (0.0025)	0.0001 (0.0006)	− 0.0045*** (0.0001)
ln(Avg h index)	0.0032* (0.0013)	− 0.0072*** (0.0013)	9.55e−5 (0.0003)	− 0.0005*** (6.33e−5)
Team hierarchy	− 0.0054 (0.0039)	− 0.0034 (0.0040)	0.0010 (0.0009)	− 0.0013*** (0.0002)
ln(Avg career age)	− 0.0007 (0.0019)	0.0056** (0.0020)	− 0.0004 (0.0004)	− 0.0011*** (9.16e−5)
Subfield FE	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes
Journal Q FE	Yes	Yes	Yes	Yes
R^2	0.59870	0.31615	0.01925	0.10266

Robust standard errors are reported in parentheses. *$p<0.05$, **$p<0.01$, ***$p<0.001$

Table 3 Effect of the data-based papers on the interdisciplinarity of papers (OLS regression)

Models	(1)	(2)	(3)	(4)
Field/data	Social science (MAG)			
D.V	ln(Variety)	Entropy	ln(Disparity)	Rao-Stirling
Data-based papers	0.0442*** (0.0027)	0.0498*** (0.0032)	0.0119*** (0.0007)	0.0079*** (0.0005)
ln(#Ref)	0.6076*** (0.0026)	0.3906*** (0.0030)	0.0633*** (0.0009)	0.0214*** (0.0005)
ln(Team size)	0.0557*** (0.0032)	0.0501*** (0.0039)	0.0069*** (0.0008)	0.0055*** (0.0006)
Gender diversity	0.0164** (0.0051)	0.0183** (0.0061)	0.0029* (0.0012)	0.0026** (0.0009)
P (female)	− 0.0273*** (0.0040)	− 0.0389*** (0.0046)	− 0.0017. (0.0010)	− 0.0061*** (0.0007)
ln(Avg h index)	0.0258*** (0.0032)	− 0.0066. (0.0037)	0.0075*** (0.0008)	− 0.0018*** (0.0005)
Team hierarchy	− 0.0439*** (0.0085)	− 0.0758*** (0.0099)	0.0061** (0.0022)	− 0.0068*** (0.0015)
ln(Avg career age)	− 0.0260*** (0.0044)	0.0033 (0.0048)	− 0.0063*** (0.0012)	0.0024*** (0.0007)
Subfield FE	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes
Journal Q FE	Yes	Yes	Yes	Yes
R^2	0.59561	0.32125	0.23290	0.08023

Robust standard errors are reported in parentheses. *$p<0.05$, **$p<0.01$, ***$p<0.001$

Table 4 Effect of the data-based papers on the disruptive potentials of papers (OLS regression)

Models	(1)	(2)	(3)	(4)	(5)	(6)
Field/data	Life science (PubMed)			Social science (MAG)		
D.V	CD	$\ln(n_i)$	$\ln(n_j)$	CD	$\ln(n_i)$	$\ln(n_j)$
Data-based papers	0.0002* (7.74e−5)	0.0771*** (0.0038)	0.0204*** (0.0037)	0.0018*** (0.0002)	0.4035*** (0.0079)	0.3016*** (0.0076)
ln(#Ref)	− 0.0041*** (0.0003)	0.1646*** (0.0032)	0.4896*** (0.0030)	− 0.0136*** (0.0006)	0.2041*** (0.0065)	0.7214*** (0.0060)
ln(Team size)	− 0.0014*** (8.68e−5)	0.3284*** (0.0042)	0.4683*** (0.0041)	− 0.0006* (0.0003)	0.2020*** (0.0101)	0.2410*** (0.0095)
Gender diversity	− 0.0004 (0.0003)	− 0.0811*** (0.0100)	− 0.1168*** (0.0098)	0.0010* (0.0005)	0.0074 (0.0150)	− 0.0261. (0.0144)
P (female)	0.0008** (0.0002)	− 0.0279** (0.0087)	− 0.0906*** (0.0086)	0.0016*** (0.0003)	0.1320*** (0.0110)	− 0.0272* (0.0108)
ln(Avg h index)	− 0.0007*** (0.0001)	0.3410*** (0.0050)	0.4936*** (0.0049)	7.4e−5 (0.0003)	0.2810*** (0.0090)	0.3563*** (0.0087)
Team hierarchy	0.0005 (0.0005)	− 0.2134*** (0.0138)	− 0.2453*** (0.0137)	− 0.0020* (0.0008)	− 0.0393 (0.0249)	− 0.0433. (0.0239)
ln(Avg career age)	− 0.0007** (0.0002)	− 0.3904*** (0.0066)	− 0.3848*** (0.0065)	0.0002 (0.0005)	− 0.1123*** (0.0115)	− 0.1765*** (0.0112)
Subfield FE	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Journal Q FE	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.03209	0.27376	0.40869	0.09901	0.34913	0.43342

Robust standard errors are reported in parentheses. *$p<0.05$, **$p<0.01$, ***$p<0.001$

Table 5 Descriptive statistics of data-based papers and the control group

	Data-based papers				Control group			
	Obs	Mean	Std	Median	Obs	Mean	Std	Median
Life science (PubMed)								
Citations till 2024	162,542	46.20	235.22	11.00	171,593	23.47	72.62	10.00
Citations (5 year)	162,542	30.34	122.55	9.00	171,593	16.43	34.21	8.00
Variety	142,434	49.73	14.52	49.00	161,133	48.62	15.54	48.00
Entropy	142,434	3.36	0.28	3.38	161,133	3.32	0.30	3.34
Disparity	142,241	0.53	0.09	0.54	161,132	0.53	0.09	0.54
Rao-Stirling	133,635	0.065	0.01	0.07	149,124	0.063	0.01	0.06
CD	158,026	0.001	0.02	0.00	169,576	0.000	0.02	0.00
n_i	180,352	15.03	144.69	2.00	182,620	7.05	49.33	2.00
n_j	180,352	26.66	114.59	6.00	182,620	15.47	34.75	6.00
Social science (MAG)								
Citations till 2021	42,018	73.38	449.10	25.00	39,851	46.04	149.51	18.00
Citations (5 year)	42,018	20.41	60.30	10.00	39,851	17.07	33.89	8.00
Variety	37,606	9.12	4.31	9.00	37,440	9.08	4.94	8.00
Entropy	37,606	1.75	0.50	1.80	37,440	1.71	0.55	1.75
Disparity	37,606	0.90	0.15	0.95	37,440	0.89	0.18	0.95
Rao-Stirling	37,606	0.23	0.06	0.23	37,440	0.22	0.07	0.22
CD	38,989	0.003	0.04	0.00	37,786	0.001	0.03	0.00
n_i	40,906	33.11	410.93	8.00	40,605	16.19	78.62	5.00
n_j	40,906	38.47	96.89	14.00	40,605	28.08	79.54	10.00

Appendix B: Figure

See Fig. 8.

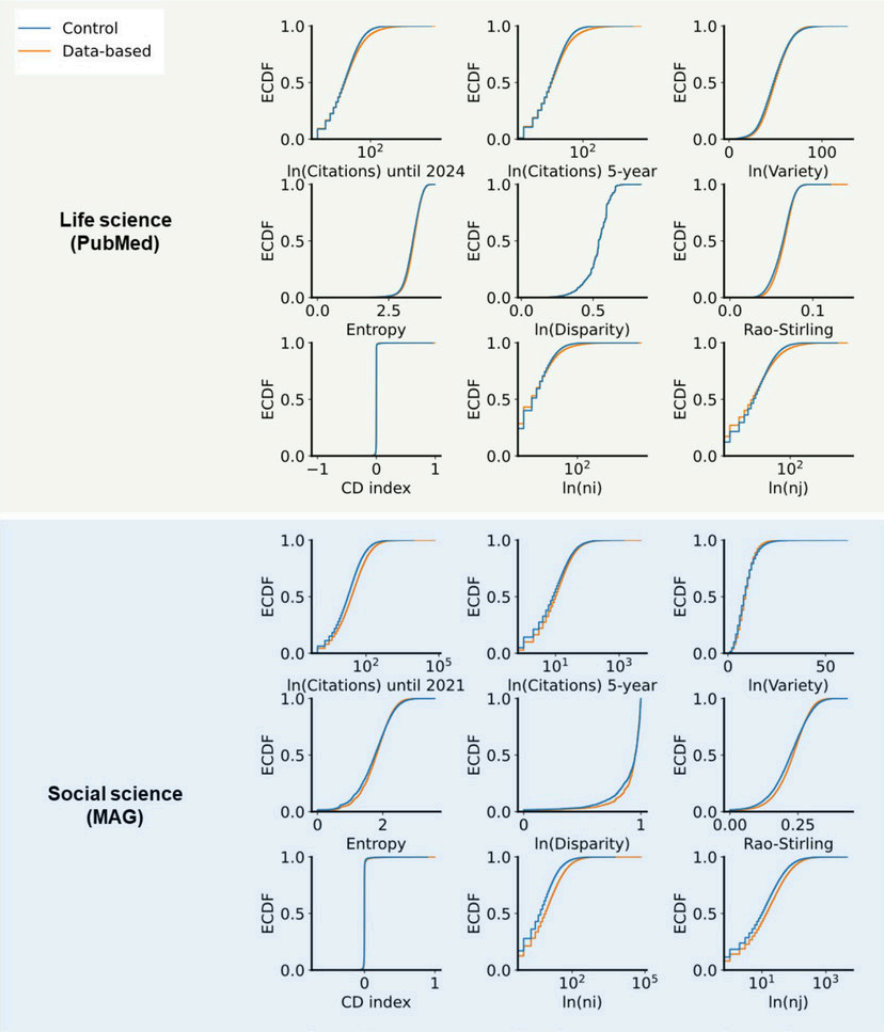


Fig. 8 Empirical cumulative distribution functions (ECDFs) of innovation measures (long-term citations, 5-year citations, variety, entropy, disparity, Rao-Stirling index, CD index, disruptive citations (ni), consolidating citations (nj)) for data-based papers and control group papers in life sciences (PubMed) and social sciences (MAG)

Appendix C: Computational complexity

To provide insight into the computational demands of our analyses, we report the approximate time required to perform key procedures, conducted on a high-performance computing cluster with 64 cores (2.5 GHz) and 256 GB RAM. The identification of data-based papers (Sect. "Data-based papers") took approximately 6 h, involving dataset mapping for 247,847 life sciences papers and 57,913 social sciences papers. The Coarsened Exact Matching (CEM) procedure (Sect. "Coarsened exact matching") required approximately 96 h for matching 189,557 life sciences papers and 24 h for matching

45,244 social sciences papers, including iterative adjustments and manual verification of control group papers to ensure accurate classification. Calculating interdisciplinarity metrics (variety, entropy, disparity, Rao-Stirling) across reference distributions for all matched papers took around 24 h for the PubMed dataset (approximately 30 million papers) and 40 h for the Microsoft Academic Graph (MAG) dataset (approximately 200 million papers). Disruption metrics, including the CD index and counts of disruptive and consolidating citations, required approximately 48 h for the PubMed dataset and 60 h for the MAG dataset, due to the computational intensity of processing citation networks. Multivariable OLS regressions, including fixed effects for publication year, subfield, and journal impact quantile, were relatively efficient, taking approximately 2 h per discipline.

Acknowledgements This research was supported by the National Social Science Fund of China (23&ZD224).

Author contributions Alex J. Yang: Writing-original draft, Conceptualization, Methodology, Software, Data visualization, Validation, Data curation, Resources. Star X. Zhao: Conceptualization, Methodology, Writing-review & editing, Supervision, Funding acquisition. Sanhong Deng: Writing- review & editing, Supervision.

Declarations

Conflict of interest We declare no competing interest relevant to this article.

References

Acuna, D. E., Yi, Z., Liang, L., & Zhuang, H. (2022). Predicting the usage of scientific datasets based on article, author, institution, and journal bibliometrics. *iConference*,

Buneman, P., Dosso, D., Lissandrini, M., & Silvello, G. (2021). Data citation and the citation graph. *Quantitative Science Studies, 2*(4), 1399–1422. https://doi.org/10.1162/qss_a_00166

Candela, L., Castelli, D., Manghi, P., & Tani, A. (2015). Data journals: A survey. *Journal of the Association for Information Science and Technology, 66*(9), 1747–1762. https://doi.org/10.1002/asi.23358

Chao, T. C. (2011). Disciplinary reach: Investigating the impact of dataset reuse in the earth sciences. *Proceedings of the American Society for Information Science and Technology, 48*(1), 1–8. https://doi.org/10.1002/meet.2011.14504801125

Du, C., Cohoon, J., Lopez, P., & Howison, J. (2021). Softcite dataset: A dataset of software mentions in biomedical and economic research publications. *Journal of the Association for Information Science and Technology, 72*(7), 870–884. https://doi.org/10.1002/asi.24454

Fortunato, S., Bergstrom, C. T., Boerner, K., Evans, J. A., Helbing, D., Milojevic, S., Petersen, A. M., Radicchi, F., Sinatra, R., Uzzi, B., Vespignani, A., Waltman, L., Wang, D., & Barabasi, A.-L. (2018). Science of science. *Science, 359*(6379), eaao0185. https://doi.org/10.1126/science.aao0185

Funk, R. J., & Owen-Smith, J. (2017). A dynamic network measure of technological change. *Management Science, 63*(3), 791–817. https://doi.org/10.1287/mnsc.2015.2366

Furman, J. L., & Stern, S. (2011). Climbing atop the shoulders of giants: The impact of institutions on cumulative research. *American Economic Review, 101*(5), 1933–1963. https://doi.org/10.1257/aer.101.5.1933

Gates, A. J., Ke, Q., Varol, O., & Barabási, A. L. (2019). Nature's reach: Narrow work has broad impact. *Nature, 575*(7781), 32–34. https://doi.org/10.1038/d41586-019-03308-7

He, L., & Han, Z. (2017). Do usage counts of scientific data make sense? An investigation of the Dryad repository. *Library Hi Tech, 35*(2), 332–342. https://doi.org/10.1108/LHT-12-2016-0158

Hemphill, L., Thomer, A., Lafia, S., Fan, L., Bleckley, D., & Moss, E. (2024). A dataset for measuring the impact of research data and their curation. *Scientific Data, 11*(1), 442. https://doi.org/10.1038/s41597-024-03303-2

Iacus, S. M., King, G., & Porro, G. (2012). Causal inference without balance checking: Coarsened exact matching. *Political Analysis, 20*(1), 1–24. https://doi.org/10.1093/pan/mpr013

Kedrick, K., Levitskaya, E., & Funk, R. J. (2024). Conceptual structure and the growth of scientific knowledge. *Nature Human Behaviour, 8*(10), 1915–1923. https://doi.org/10.1038/s41562-024-01957-x

Koesten, L., Vougiouklis, P., Simperl, E., & Groth, P. (2020). Dataset reuse: Toward translating principles to practice. *Patterns, 1*(8), Article 100136. https://doi.org/10.1016/j.patter.2020.100136

Leydesdorff, L., Wagner, C. S., & Bornmann, L. (2019). Interdisciplinarity as diversity in citation patterns among journals: Rao-stirling diversity, relative variety, and the gini coefficient. *Journal of Informetrics, 13*(1), 255–269. https://doi.org/10.1016/j.joi.2018.12.006

Lin, Z., Yin, Y., Liu, L., & Wang, D. (2023). SciSciNet: A large-scale open data lake for the science of science research. *Scientific Data, 10*(1), 315. https://doi.org/10.1038/s41597-023-02198-9

Liu, L., Jones, B. F., Uzzi, B., & Wang, D. (2023a). Data, measurement and empirical methods in the science of science. *Nature Human Behaviour*. https://doi.org/10.1038/s41562-023-01562-4

Liu, L., Jones, B. F., Uzzi, B., & Wang, D. (2023b). Data, measurement and empirical methods in the science of science. *Nature Human Behaviour, 7*(7), 1046–1058. https://doi.org/10.1038/s41562-023-01562-4

Park, M., Leahey, E., & Funk, R. J. (2023). Papers and patents are becoming less disruptive over time. *Nature, 613*(7942), 138–144. https://doi.org/10.1038/s41586-022-05543-x

Qin, J., Hemsley, J., & Bratt, S. E. (2022). The structural shift and collaboration capacity in GenBank networks: A longitudinal study. *Quantitative Science Studies, 3*(1), 174–193. https://doi.org/10.1162/qss_a_00181

Robinson-Garcia, N., Mongeon, P., Jeng, W., & Costas, R. (2017). Datacite as a novel bibliometric source: Coverage, strengths and limitations. *Journal of Informetrics, 11*(3), 841–854. https://doi.org/10.1016/j.joi.2017.07.003

Stirling, A. (2007). A general framework for analysing diversity in science, technology and society. *Journal of the Royal Society Interface, 4*(15), 707–719. https://doi.org/10.1098/rsif.2007.0213

Sung, M., Jeong, M., Choi, Y., Kim, D., Lee, J., & Kang, J. (2022). BERN2: An advanced neural biomedical named entity recognition and normalization tool. *Bioinformatics, 38*(20), 4837–4839. https://doi.org/10.1093/bioinformatics/btac598

Wang, D., & Barabási, A.-L. (2021). *The science of science*. Cambridge University Press.

Wang, K., Shen, Z., Huang, C., Wu, C.-H., Dong, Y., & Kanakia, A. (2020). Microsoft academic graph: When experts are not enough. *Quantitative Science Studies, 1*(1), 396–413. https://doi.org/10.1162/qss_a_00021

Wu, L. F., Wang, D. S., & Evans, J. A. (2019). Large teams develop and small teams disrupt science and technology. *Nature, 566*(7744), 378. https://doi.org/10.1038/s41586-019-0941-9

Xu, J., Yu, C., Xu, J., Ding, Y., Torvik, V. I., Kang, J., Sung, M., & Song, M. (2024). PubMed knowledge graph 2.0: Connecting papers, patents, and clinical trials in biomedical science. *preprint, arXiv 2410.07969*.

Xu, F., Wu, L., & Evans, J. (2022). Flat teams drive scientific innovation. *Proceedings of the National Academy of Sciences of the United States of America, 119*(23), Article e2200927119. https://doi.org/10.1073/pnas.2200927119

Yang, A. J. (2025). Understanding the relationship between interdisciplinary knowledge and disruption in science. *Journal of Information Science*. https://doi.org/10.1177/01655515251330614

Yang, A. J., Gong, H., Wang, Y., Zhang, C., & Deng, S. (2024). Rescaling the disruption index reveals the universality of disruption distributions in science. *Scientometrics, 129*(1), 561–580. https://doi.org/10.1007/s11192-023-04889-x

Yang, A. J., Zhao, S. X., Deng, S., Liu, M., Bu, Y., & Ding, Y. (2025). Quantifying the dynamics of research teams' academic diversity. *Journal of the Association for Information Science and Technology,* 1–17. https://doi.org/10.1002/asi.70023

Yang, A. J., Wang, F., Shi, Y., Zhang, Y., Wang, H., & Deng, S. (2025a). Beyond surface correlations: Reference behavior mediates the disruptiveness-citation relationship. *Journal of Data and Information Science*. https://doi.org/10.2478/jdis-2025-0029

Yang, A. J., Yan, X., Hu, H., Hu, H., Kong, J., & Deng, S. (2025b). Are disruptive papers more likely to impact technology and society? *Journal of the Association for Information Science and Technology, 76*(3), 563–579. https://doi.org/10.1002/asi.24947

Yang, Y., Tian, T. Y., Woodruff, T. K., Jones, B. F., & Uzzi, B. (2022). Gender-diverse teams produce more novel and higher-impact scientific ideas. *Proceedings of the National Academy of Sciences of the United States of America, 119*(36), Article e2200841119. https://doi.org/10.1073/pnas.2200841119

Yin, Y., Dong, Y., Wang, K., Wang, D., & Jones, B. F. (2022). Public use and public funding of science. *Nature Human Behaviour, 6*(10), 1344–1350. https://doi.org/10.1038/s41562-022-01397-5

Yin, Y., Gao, J., Jones, B. F., & Wang, D. (2021). Coevolution of policy and science during the pandemic. *Science, 371*(6525), 128–130. https://doi.org/10.1126/science.abe3084

Youyou, W., Yang, Y., & Uzzi, B. (2023). A discipline-wide investigation of the replicability of psychology papers over the past two decades. *Proceedings of the National Academy of Sciences, 120*(6), Article e2208863120. https://doi.org/10.1073/pnas.2208863120

Yu, Y., & Romero, D. M. (2024). Does the use of unusual combinations of datasets contribute to greater scientific impact? *Proceedings of the National Academy of Sciences of the United States of America, 121*(41), Article e2402802121. https://doi.org/10.1073/pnas.2402802121

Zanin, M., Papo, D., Sousa, P. A., Menasalvas, E., Nicchi, A., Kubik, E., & Boccaletti, S. (2016). Combining complex networks and data mining: Why and how. *Physics Reports, 635*, 1–44.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.