An analysis of the effects of open science indicators on citations in the French Open Science Monitor

Giovanni Colavizza†
University of Copenhagen, Denmark
University of Bologna, Italy
Odoma LLC, Switzerland

Lauren Cadwallader
PLOS, United States

Iain Hrynaszkiewicz††
PLOS, United States

August 29, 2025

Abstract

This study investigates the correlation of citation impact with various open science indicators (OSI) within the French Open Science Monitor (FOSM), a dataset comprising approximately 900,000 publications authored by French authors from 2020 to 2022. By integrating data from OpenAlex and Crossref, we analyze open science indicators such as the presence of a pre-print, data sharing, and software sharing in 576,537 publications in the FOSM dataset. Our analysis reveals a positive correlation between these OSI and citation counts. Considering our most complete citation prediction model, we find pre-prints are correlated with a significant positive effect of 19% on citation counts, software sharing of 13.5%, and data sharing of 14.3%. We find large variations in the correlations of OSIs with citations in different research disciplines, and observe that open access status of publications is correlated with a 8.6% increase in citations in our model. While these results remain observational and are limited to the scope of the analysis, they suggest a consistent correlation between citation advantages and open science indicators. Our results may be valuable to policy makers, funding agencies, researchers, publishers, institutions, and other stakeholders who are interested in understanding the academic impacts, or effects, of open science practices.

1 Introduction

The French Open Science Monitor (FOSM; `https://frenchopensciencemonitor.esr.gouv.fr`) is an initiative of the French Ministry of Higher Education and Research and more specifically the French Committee for Open Science, which ensures implementation of the French National Open Science Policy. To enable the French Committee for Open Science to monitor (measure) the impact of its National Open Science Policy, it introduced the FOSM in 2018, to measure the progress in the adoption of open science practices in France. The first edition of the FOSM in 2018 provided indicators on open access to scientific publications. The 2022 edition introduced indicators on sharing of research data, and sharing of code and software [1]. FOSM is based on open data sources and uses open source tools to produce its results. These include (meta)data from Unpaywall, the Directory of Open Access Journals (DOAJ), PubMed, Crossref, article PDFs and web crawling techniques. The scope of the FOSM includes all research publications that can be discovered using its methodology with at least one French author. The software tools used to produce the FOSM are open source and include GROBID, a tool for extracting structured data (in XML format) from PDFs; DataStet, a tool that identifies mentions of research data and research datasets in publications; and Softcite, which identifies code and software mentions in publications (`https://barometredelascienceouverte.esr.gouv.fr/about/methodology`). The data produced by the FOSM to create its visualizations and indicators are also made available under an open licence (`https://barometredelascienceouverte.esr.gouv.fr/about/opendata`), and institutions, such as the University of Lorraine, have adapted the Monitor to create local versions.

†colavizza@hum.ku.dk
††ihrynaszkiewicz@plos.org

The FOSM – and initiatives from members of our group such as PLOS Open Science Indicators [2] – is part of a growing number of global initiatives focused on open science monitoring. Open science monitors have so far largely focused on monitoring open science practices and outputs, such as publications and research data, but open science monitoring applies to all aspects of open science and its effects, as defined by the 2021 UNESCO definition on open science. [3] Indeed, UNESCO is a partner of the Open Science Monitoring Initiative (OSMI; `https://open-science-monitoring.org`), which launched in 2024 and aims to promote the adoption of common principles [4] and frameworks for open science monitoring globally, along with specifications for their implementation. With this broad scope of open science monitoring, measuring the outcomes (or impacts/consequences) as well as the processes of open science – across diverse regions, contexts, and disciplines – are equally important to monitoring the outputs of open science practices, [5] . However, outcomes and impacts of open science are generally more challenging to measure at a large scale. Developing better approaches to measuring the impacts of open science is a focus of the PathOS (Open Science Impact Pathways; `https://pathos-project.eu`) project, which has produced an Open Science Indicators handbook. [6] PathOS has also produced systematic scoping reviews on the impacts of open science in three areas: academic impacts, societal impacts, and economic impacts. [7, 8, 9]

In their systematic scoping review of academic impacts, Klebel *et al*. [8] identified citations, quality, efficiency, equity, reuse, ethics and reproducibility as areas of academic impact that have been studied. For our current study, the academic impacts of open science are most relevant. Although there are limitations to what citations can tell evaluators and users of research about quality and impact, citations remain an important metric for understanding the impact of research findings. Due to their widespread use, citations can be measured across large corpora of articles. Klebel *et al.* [8] identified 22 studies reporting an impact of open data (data sharing) on citations; four reporting a citation advantage for pre-prints; and three finding a citation advantage for open code (code sharing). However, none confirms a causal relationship between open science practices and citation advantages. Our recent study complements these findings by finding a 20.2% citation advantage for the sharing of pre-print and 4.3% advantage for the sharing of data in a repository. [10] However, we did not find a statistically significant citation advantage for code sharing. Our previous analysis used the PLOS Open Science Indicators dataset, which is produced in collaboration with DataSeer, and, at the time of writing, has monitored five open science practices including data, code, and pre-print sharing, as well as, more recently, protocol sharing, and study registration (preregistration). As an ongoing open science monitoring initiative, the PLOS Open Science Indicators dataset is updated periodically as new PLOS and comparator content is published. Our previous study was based on version 5 of the dataset, which included 121,999 articles, the majority of which are PLOS articles. [11] Given the alignment of PLOS Open Science Indicators with the indicators included in the FOSM and its production of reusable open data, we sought to extend this previous work on measuring the citation effects of open science practices to the FOSM's larger and more diverse dataset (`https://data.enseignementsup-recherche.gouv.fr/explore/dataset/open-access-monitor-france/information`). Furthermore, given the purpose of the FOSM is to inform understanding of open science policy effects, we assumed that information on the impact of open science practices – in addition to information on the prevalence of those practices that is included in the FOSM – will be valuable for policy makers, funding agencies, researchers, publishers, institutions, and other stakeholders.

2 Data and Methods

The 2023 edition of the FOSM dataset contained approximately 900'000 publications at the time of our analysis. The indicators of interest include whether data and software were created and shared, and whether there are pre-prints associated with published articles. We seek to answer the research question based on the FOSM dataset, namely, whether and to what extent a citation impact premium is received on average by articles following some or all of the open science practices under consideration. This work follows the methodology, and expands upon the results of a previously published work [12, 10], including PLOS' Open Science Indicators [11].

To collect the data for analysis, we proceed as follows:

1. Acquire the FOSM dataset (`https://data.enseignementsup-recherche.gouv.fr/explore/dataset/open-access-monitor-france/information`). We use the December 30, 2022 snapshot (end of coverage period). This snapshot includes 897'426 entries. Of these, 545'158 are classified as journal articles (60.75%). The publications were published between 2020 and 2022, both years included. The number of publications per year is roughly uniform.

2. Query OpenAlex (`https://openalex.org`). for every DOI present in FOSM, and download the full OpenAlex record locally. OpenAlex was queried between July 12 and 15, 2024. The number of entries in FOSM with an OpenAlex match, after the removal of duplicates, is 576'537 (64.24%). This dataset contains 479'700 journal articles (83.2% of the matched total), and is used in what follows.

3. Add pre-print information from Crossref, using an experimental pre-print-publication relationship dataset created by Crossref. [13] This dataset covers Crossref publications up to the end of August 2023, therefore fully matching the FOSM coverage. The number of entries in this Crossref dataset is 641'950. Of these, 22'303 match with both FOSM and OpenAlex (3.9% of FOSM with an OpenAlex match). The total number of pre-print matches, including also arXiv matches indexed in OpenAlex, is 44'763 (7.8%). This proportion follows known trends in the literature [14].

From OpenAlex, for every entry, we take the month of publication, the number of authors and references, whether there is a pre-print in arXiv, and the citation count at the time of the query (July 2024). From Crossref, we recover possible further matches with other pre-print servers. From FOSM, we take every other variable we use in what follows.

We provide several statistics on the available variables. Table 1 shows the publication counts for different BSO categories. BSO (Baromètre de la Science Ouverte) is a system to classify scientific domains into macro areas. In FOSM, each publication is assigned a BSO class automatically, see [15] for methodological details. Next, we show the publication counts by publication typology (genre) (Table 2), showing a large majority for journal articles, by access status (Table 3) and open access status (Table 4). Lastly, we provide descriptive statistics for the main dependent variable (citation counts), controls (Table 5), and open science indicator controls (Table 6). We note that the maximum correlation among controls is 0.25 for software and data sharing, thus ruling out issues due to multicollinearity. This is part of standard checks done before model fitting, that include the need to avoid using highly correlated controls which would make the model fitting less reliable.

From now on, we consider as open science indicators (OSI) the following variables of interest: pre-print publication, software shared, data shared. We also include software used and created, and data used and created. Although strictly speaking these are not open science practices, they are useful comparisons for sharing activities.

Table 1: Publication counts by BSO category.

BSO category	Publication counts
Medical research	160'672
Biology (fond.)	95'019
Earth, Ecology, Energy and applied biology	55'852
Humanities	51'309
Physical sciences, Astronomy	48'659
Social sciences	43'164
Computer and information sciences	37'948
Engineering	36'516
Chemistry	28'856
Mathematics	18'437
unknown	105

Table 2: Publication counts by typology (genre).

Publication genre	Publication counts
Journal article	470'434
Book chapter	45'795
Proceedings	29'191
Other	18'127
pre-print	9'636
Book	3354

Table 3: Publication counts by access status.

Access status	Publication counts
Closed	193'222
Publisher-repository	183'554
Repository	102'907
Publisher	96'534

Table 4: Publication counts by open access status.

Open access status	Publication counts
Gold	132'435
Green	107'370
Hybrid	92'824
Bronze	51'329

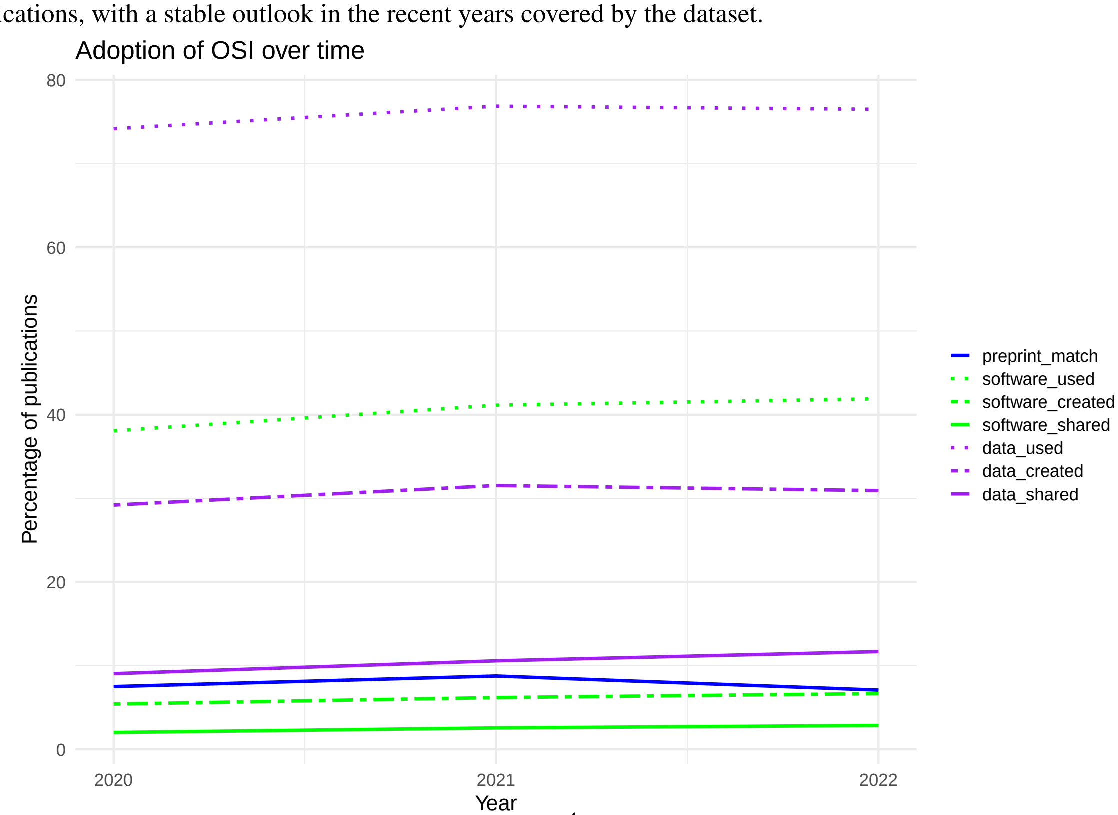
Table 5: Descriptive statistics for the dependent variable and a set of publication and author level controls.

Statistic	Cited by Count	Number of Authors	Number of References	Year	Month
Min.	0	0	0	2020	1
1st Qu.	0	2	1	2020	3
Median	2	5	23	2021	6
Mean	11	6.9	32	2021	6
3rd Qu.	10	8	46	2022	9
Max.	63705	100	4083	2022	12

Table 6: Descriptive statistics for open science indicators.

Variable	FALSE Count	TRUE Count	NA Count
pre-print match	531'774	44'763	0
Software used	202'336	136'300	237'901
Software created	318'136	20'500	237'901
Software shared	330'305	8331	237'901
Data used	81'890	256'201	238'446
Data created	234'950	103'141	238'446
Data shared	303'080	35'011	238'446
Is OA	193'222	382'995	320

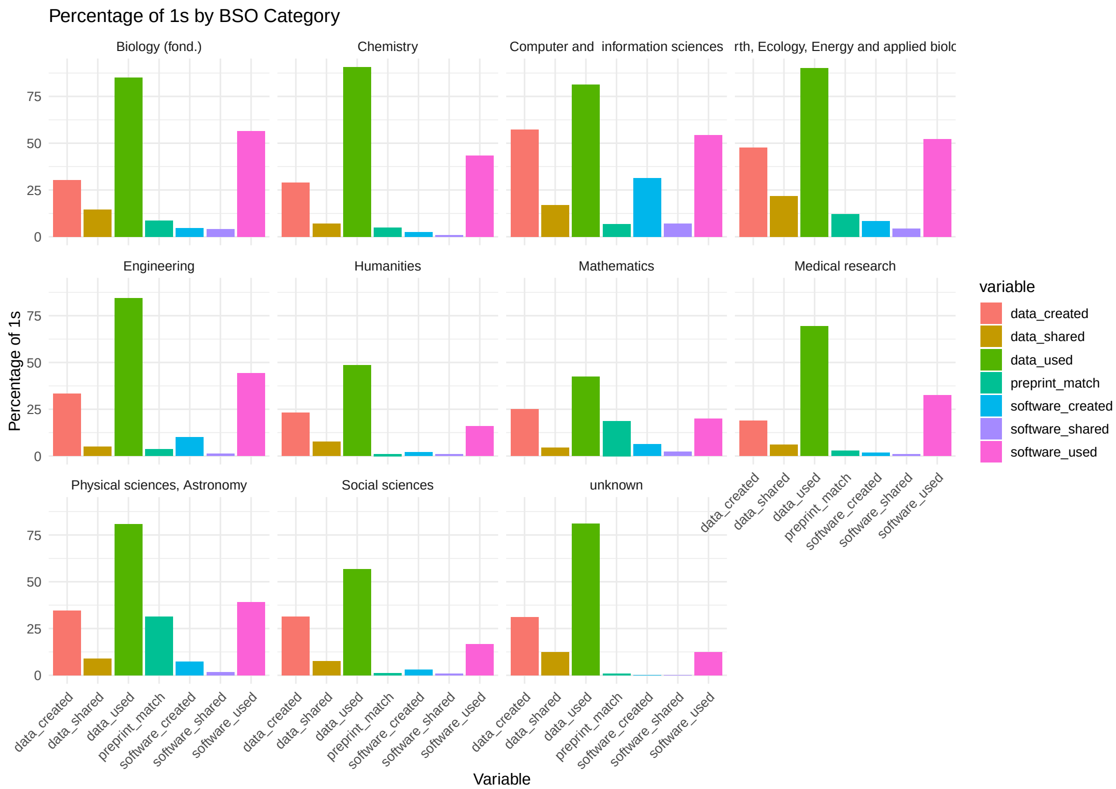
Figure 1: Adoption of open science indicators (OSI) over time in FOSM. Each OSI remains adopted by a fraction of publications, with a stable outlook in the recent years covered by the dataset.



3 Results

We start by visualizing the temporal trends in OSI adoption, reminding the reader that the FOSM dataset used in our analysis only covers publication from 2020, 2021, and 2022. In Figure 1, we notice how most OSI remain stable and relatively low over time, despite significant levels of data and software use. In Figure 2, we unpack trends by BSO class, noticing some common trends such as higher software creation in Computer Science, and pre-print adoption in Physics and Mathematics.

Figure 2: Adoption of OSI by BSO class, as shown in Table 1. Each OSI remains adopted by a fraction of publications, and there is significant variation across domains.



3.1 Modeling

The base model we use is described in Equation 1, and the full model is described in Equation 2. Variable transformations are shown, numerical variables are given in Italics, and categorical variables are in regular text. Variables are grouped along lines. This model follows closely from our previous work, where an illustration of the assumed causal dependency graph among variable groups is also provided [10]. We notice that, with respect to our previous work, we have no ability to control for the mean h-index of the authors at the time of publication, and we use the BSO classification instead of the Fields of Research classification codes (`https://www.arc.gov.au/manage-your-grant/classification-codes-rfcd-seo-and-anzsic-codes`). At the same time, we add further variables such as software/data used or created, publication genre (e.g. Journal article, Book chapter), and open access status. The absence of the h-index control is the most significant limitation of our work here, in terms of its comparability to our previous work, as it constitutes a proxy for author success or reputation at the time of publication, which might also correlate with resource availability. To confirm the absence of discrepancies, we provide results on a small set of journal articles that are part of both PLOS OSI and FOSM in the Appendix.

$$\begin{aligned}\log(n\_cit\_tot + 1) = &\log(n\_authors + 1) + \log(n\_references + 1) + year + month + \\ &\text{pre-print\_match} + \\ &\text{software\_created} + \text{software\_shared} + \\ &\text{data\_created} + \text{data\_shared}\end{aligned} \tag{1}$$

$$
\begin{aligned}
\log(n\_cit\_tot + 1) = & \log(n\_authors + 1) + \log(n\_references + 1) + year + month + \\
& \text{genre} + \text{pre-print\_match} + \\
& \text{software\_used} + \text{software\_created} + \text{software\_shared} + \\
& \text{data\_used} + \text{data\_created} + \text{data\_shared} + \\
& \text{is\_oa} + \text{bso\_classification}
\end{aligned} \quad (2)
$$

Starting with the base model in Table 7, we provide results for an Ordinary Least Squares (OLS) model and a robust linear model as a comparison. The results are aligned and show a relatively high explained variance with the base model having $R^2 = .45$. The model shows expected trends with respect to common controls (authors, references, year of publication), and interesting novel trends with respect to OSI. The OSI show that there is a significant and positive correlation of pre-prints (22.6%), sharing data (14.1%) and sharing software (6.5%) with increased citations. These percentage changes for log-linear relationships are calculated as follows: $(\exp(.204) - 1) \times 100 \approx 22.6\%$. Our next question is whether these results hold when we account for large disciplinary variations in the adoption of open science practices, as well as further controls which we add next.

Table 7: Results

	Dependent variable:	
	n_cit_tot_log	
	OLS	*robust linear*
	(1)	(2)
n_authors_log	0.399^{***} (0.002)	0.386^{***} (0.002)
n_references_log	0.454^{***} (0.001)	0.468^{***} (0.001)
year	-0.330^{***} (0.002)	-0.324^{***} (0.002)
month	-0.023^{***} (0.0005)	-0.023^{***} (0.0004)
C(pre-print_match)	0.204^{***} (0.005)	0.195^{***} (0.005)
C(software_created)	0.028^{***} (0.007)	0.029^{***} (0.007)
C(software_shared)	0.063^{***} (0.011)	0.063^{***} (0.011)
C(data_created)	0.127^{***} (0.004)	0.122^{***} (0.004)
C(data_shared)	0.132^{***} (0.006)	0.127^{***} (0.006)
Constant	666.250^{***} (4.173)	653.919^{***} (4.081)
Observations	337,929	337,929
R^2	0.450	
Adjusted R^2	0.450	
Residual Std. Error (df = 337919)	0.970	0.880

F Statistic	30,682.610*** (df = 9; 337919)
Note:	*p<0.1; **p<0.05; ***p<0.01

In the full model we propose, in Table 8, we add the following variables: publication genre, software/data used, open access status, language, and BSO classification. We take the categories with most publications as reference categories, specifically: 'journal-article' for genre, 'English' for language, and 'Medical research' for BSO classification. Our model confirms previous trends, namely pre-print match contributing to a significant and positive correlation of 19% with citation counts, software sharing of 13.5%, and data sharing of 14.3%. We also find positive effects from software/data creation. Publications where data is used tend to be more cited, whilst those when software is used slightly less so. Publications in English are significantly more cited than publications in other languages. Disciplinary differences (with respect to the baseline Medical research), and publication genre differences (with respect to the baseline journal article), also follow known trends.

While open access (OA) status is not a relevant indicator for the PLOS OSI dataset given all content is OA, OA status is an important aspect of the FOSM, with OA adoption reaching 65.6% in 2022 in French publications (https://frenchopensciencemonitor.esr.gouv.fr/). We included OA status in our model and can observe a 8.6% citation increase correlated with publications that are OA compared to those that are closed access. For a review on the citation impact of open access status, see [16]. OA status can be explored in more depth considering OA typologies using our data and code. [17]

Table 8: Results

	Dependent variable:	
	n_cit_tot_log	
	OLS	*robust linear*
	(1)	(2)
n_authors_log	0.346***	0.326***
	(0.003)	(0.002)
n_references_log	0.379***	0.382***
	(0.001)	(0.001)
year	−0.327***	−0.316***
	(0.002)	(0.002)
month	−0.023***	−0.022***
	(0.0004)	(0.0004)
C(genre)book	0.635***	0.604***
	(0.032)	(0.032)
C(genre)book-chapter	−0.686***	−0.668***
	(0.010)	(0.010)
C(genre)other	−0.541***	−0.502***
	(0.013)	(0.013)
C(genre)preprint	−1.392***	−1.423***
	(0.011)	(0.011)
C(genre)proceedings	−0.534***	−0.541***
	(0.010)	(0.010)
C(preprint_match)	0.174***	0.177***
	(0.006)	(0.005)
C(software_used)	−0.057***	−0.030***

	(0.004)	(0.004)
C(software_created)	0.061^{***}	0.053^{***}
	(0.007)	(0.007)
C(software_shared)	0.127^{***}	0.116^{***}
	(0.011)	(0.011)
C(data_used)	0.148^{***}	0.162^{***}
	(0.005)	(0.005)
C(data_created)	0.076^{***}	0.075^{***}
	(0.004)	(0.004)
C(data_shared)	0.134^{***}	0.122^{***}
	(0.006)	(0.006)
C(is_oa)	0.083^{***}	0.075^{***}
	(0.005)	(0.005)
C(lang_reduce)fr	-0.584^{***}	-0.538^{***}
	(0.007)	(0.006)
C(lang_reduce)other	-0.562^{***}	-0.515^{***}
	(0.014)	(0.014)
C(bso_classification)Biology (fond.)	0.048^{***}	0.063^{***}
	(0.005)	(0.005)
C(bso_classification)Chemistry	0.061^{***}	0.095^{***}
	(0.007)	(0.007)
C(bso_classification)Computer and information sciences	0.107^{***}	0.118^{***}
	(0.008)	(0.008)
C(bso_classification)Earth, Ecology, Energy and applied biology	0.026^{***}	0.051^{***}
	(0.006)	(0.006)
C(bso_classification)Engineering	0.066^{***}	0.099^{***}
	(0.007)	(0.007)
C(bso_classification)Humanities	0.033^{***}	0.051^{***}
	(0.008)	(0.008)
C(bso_classification)Mathematics	-0.140^{***}	-0.115^{***}
	(0.009)	(0.009)
C(bso_classification)Physical sciences, Astronomy	-0.102^{***}	-0.059^{***}
	(0.007)	(0.006)
C(bso_classification)Social sciences	0.095^{***}	0.099^{***}
	(0.008)	(0.008)
C(bso_classification)unknown	-0.259	-0.221
	(0.231)	(0.226)
Constant	660.758^{***}	638.442^{***}
	(3.973)	(3.888)

Observations	337,928	337,928
R^2	0.503	
Adjusted R^2	0.503	
Residual Std. Error (df = 337898)	0.922	0.858
F Statistic	11,779.980*** (df = 29; 337898)	
Note:		*p<0.1; **p<0.05; ***p<0.01

These results are robust to a variety of modeling changes, which we include in our accompanying repository.

Lastly, we provide a summary table with the percentage change on citation counts linked to each OSI available, dividing publications by scientific domain (BSO class) and using only the main publication genre: journal articles. These results should be more easily readable than those provided above, when interested in differences across scientific domains. We fit a full model specification for every BSO class, and provide results in Figure 3. While these results confirm the trends discussed above, they also highlight significant variations across scientific domains. For example, the open access status is positively correlated to citation counts only in Medicine and Biology, and negatively elsewhere.

Figure 3: Percentage change on citation counts linked to each OSI, divided by BSO class. We only consider journal articles to fit these models.

Percentage Changes by BSO Classification

Significant values (p ≤ 0.05) are bold

	Computer and information sciences	Biology (fond.)	Chemistry	Medical research	Physical sciences, Astronomy	Humanities	Social sciences	Earth, Ecology, Energy and applied biology	Mathematics	Engineering
C(preprint_match)TRUE	**22**	**25.3**	**14.5**	**17.7**	**24.9**	**62.5**	-3.8	1	**9.7**	**23.1**
C(software_used)TRUE	-0.2	**-19.9**	**-11.8**	**-8.2**	0.1	**25.6**	**7.5**	**-9.3**	**8.8**	**-3.2**
C(software_created)TRUE	**-5.2**	**9.3**	**12.1**	0.3	**12.4**	**17.5**	9.4	**8.8**	**8.5**	3.4
C(software_shared)TRUE	5.3	**11.4**	**33**	5.7	**21.3**	**15**	**38**	**17.9**	12.1	**16.6**
C(data_used)TRUE	**22.9**	**7.7**	-1.3	**10.3**	**4**	**11**	**17.9**	**15.7**	**10.3**	**13.7**
C(data_created)TRUE	**13.4**	**7**	**2.9**	**13.6**	**6.1**	**6.8**	**8.9**	**2.5**	**6.4**	**8.4**
C(data_shared)TRUE	**6.5**	**18.8**	**6.4**	**34.9**	**10.4**	-1.4	2.3	**12**	**12.9**	3.2
C(is_oa)TRUE	**-17.8**	**3.7**	**-5.3**	**13.5**	**-9.7**	-0.1	**-11.2**	**-11.6**	-2.8	**-17.3**

4 Discussion and Conclusions

In this work, we have explored the FOSM dataset and investigated whether three open science indicators (OSI) – data sharing, code sharing, pre-print posting – are correlated with a citation advantage received by the publications that exhibit these open science practices. To our knowledge, this is the largest scale analysis of these open science practices and citations to date, and the first such study offering a nationwide (French) perspective on this topic . We use over half a million publications from all scientific domains, and find a positive correlation of these open science practices with citation counts. This study is observational and limited by the scope of the analysis and the available data, therefore its results should be taken with caution. We are in no position to establish causality nor to control for additional, confounding effects. Significant confounding effects that we could not consider are the quality of the research, and the reputation of the authors (as measured by h-index) and of the venue of publication. A more developed discussion of these and other limitations of this approach to measuring citation effects is provided in [10]. In theory, the observed correlation between increased citations and these open science practices in our analysis are cumulative, which would suggest a substantial average increase – 46.8% – in citations where an article exhibited all three open science practices. However, such an assumption should be made with caution given the aforementioned limitations.

Our finding of a significant positive effect of pre-print posting on citations, of 19% in this study, confirms results of previous studies. [10] Our finding of a significant positive effect of software sharing of 13.5% contrasts the result of our previous study, which found no statistically significant increase in citations correlated with code (software) sharing. [10] This may be due to the larger and more diverse, with respect to research topics and disciplines, dataset used in this study compared to our previous work. However, other studies [18] have found as much as a three-fold increase in citations associated with code sharing. Our finding of a significant positive effect of data sharing of 14.3% confirms a significant positive effect observed in two previous studies, which found increases in citations correlated with data sharing in a repository of 25.3% when analysing publications in PLOS and BMC (BioMed Central) and 4.3% when analyzing publications in PLOS OSI. [12, 10] The larger effect of OSI on citations observed in the current study may be due to the larger and more diverse corpus of publications, which includes, for example, a higher proportion of content from the Humanities and Physical Sciences than these previous two studies. Also, the FOSM publication corpus includes content that is not open access, in contrast to these previous two studies. The source of publications and citation counts (OpenAlex) in the current study is different to our previous studies, which relied on PubMed Central. OpenAlex is emerging as a credible alternative to established databases for bibliometric analyses but may lag behind in the number of indexed references (and as a consequence, citations) – at least when compared to the commercial database, Scopus. [19]

We can observe some notable differences in the correlation of citations with OSIs by BSO class (discipline), in Figure 3. For example. data sharing is correlated with a 34.9% increase in citations in Medical research, and the effect of pre-printing is particularly high in Humanities (62.5% increase) publications. Public data sharing in Medical research is often more challenging given data from research on human research participants are often sensitive or require additional processing to share. If causative effects of data sharing on citations were confirmed, we could speculate that the additional effort on behalf of researchers to share high value datasets is rewarded with greater reuse and citations of their work. In the Humanities, our sample is smaller, but it is possible cultural and temporal differences in publishing and citation norms relative to other disciplines are a factor, amplifying the observed effect. Conversely, we observe a small negative correlation of data sharing (-1.4%) in Humanities.

Regarding the larger difference we observed in the citation effect of data sharing compared to our previous study finding a 4.3% positive correlation, differences in how open science practices are defined and measured between studies are also relevant. In [10], the positive correlation was observed for a specific method of data sharing, data sharing in a repository linked to the article (as opposed to sharing data in online supplementary material). There are some subtle but important differences in how open science practices are defined and, consequently, measured, in different open science monitoring initiatives and software tools. Systematically mapping OSI definitions and measures across multiple studies is outside of the scope of the current work, although we provide a small-scale mapping of a subset of publications between our current and previous study in the Appendix. This issue highlights the need for development of common frameworks for, and consensus on, methods for open science monitoring, using open methods and data.

With the above limitations, this study adds to the evidence that OSI (open science practices) are associated with increased academic impact of research through citations. For a more comprehensive review of the academic impacts of open science practices see [8]. Future research could seek to understand if open science practices lead to other (non-citation based) academic impacts of research, such as whether they lead to increased (re)use of research outputs, improved reproducibility, or greater trust in research outputs. Much more work on determining causal mechanisms of changes in academic impacts of open science is also needed.

Data and Code Availability

Please find all data and code to reproduce our results at `https://doi.org/10.6084/m9.figshare.27822663`. [17]

Acknowledgments

For their support in using the French Open Science Monitor dataset, we thank Eric Jeangirard and Anne L'Hote at the French Ministry of Higher Education and Research, Laetitia Bracco at the University of Lorraine and Laurent Romary at the National Institute for Research in Digital Science and Technology (Inria). We also thank Marin Dacos and Arianna Caporali at the French Ministry of Higher Education and Research for participating in discussions with PLOS that contributed to the conceptualization of this research.

Author contributions

GC: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Supervision, Writing – original draft, Writing – review & editing.
LC: Formal Analysis, Visualization, Writing – review & editing.
IH: Conceptualization, Funding acquisition, Methodology, Project administration, Resources, Supervision, Writing – original draft, Writing – review & editing.

Competing interests

Two of the authors (LC and IH) were at the time of publication employed by PLOS, which produces PLOS Open Science Indicators. PLOS (IH) is a founding member and initiator of the Open Science Monitoring Initiative (OSMI). GC declares no competing interests.

Funding information

PLOS provided funding for data acquisition, modeling, and analysis, and had a role in the design, analysis, and preparation of the study. PLOS also provided support in the form of salaries for authors LC and IH.

Appendix

We report here on an analysis using the same models but focused on the overlap between FOSM and the PLOS OSI dataset (version 7). The overlap consists of 3120 publications, from PLOS and the control group. Note that some control variables (genre, open access status) had to be omitted due to missing factors. The results align with the findings of [10] and with the findings of this analysis, showing larger effects for data and software sharing.

Please note there are some differences in definitions and measures in the FOSM compared to PLOS OSI, reflecting the different methods used by the FOSM, and PLOS/DataSeer teams. For reference, the FOSM uses the following definitions for data/software use, creation, and sharing [1]:

- Data/software use refers to publications that mention the usage of at least one dataset/software.
- Data/software created refers to publications that mention the usage and the production of at least one of their datasets/software.
- Data/software shared refers to publications that mention the usage, the production and the sharing of at least one of their datasets/software.

While our measures of code sharing are well aligned across FOSM and PLOS OSI, in our previous study of PLOS OSI datasets, for data sharing, we assessed the effects of sharing research data on citations in a more specific way, that is, sharing via an online data repository [10].

Table 9: Base model, PLOS OSI and FOSM overlap.

	Dependent variable:	
	n_cit_tot_log	
	OLS	*robust linear*
	(1)	(2)
n_authors_log	0.433***	0.399***
	(0.027)	(0.027)
n_references_log	0.333***	0.335***
	(0.031)	(0.031)
year	−0.451***	−0.460***
	(0.020)	(0.019)
month	−0.039***	−0.040***
	(0.005)	(0.004)
C(pre-print_match)	0.254***	0.253***
	(0.036)	(0.035)
C(software_created)	−0.069	−0.043
	(0.064)	(0.062)

C(software_shared)	0.183***	0.163**
	(0.068)	(0.066)
C(data_created)	0.066**	0.055*
	(0.033)	(0.032)
C(data_shared)	0.158***	0.158***
	(0.047)	(0.045)
Constant	912.210***	928.924***
	(39.459)	(38.540)
Observations	2,812	2,812
R^2	0.269	
Adjusted R^2	0.267	
Residual Std. Error (df = 2802)	0.832	0.747
F Statistic	114.584*** (df = 9; 2802)	
Note:		*p<0.1; **p<0.05; ***p<0.01

Table 10: Full model, PLOS OSI and FOSM overlap.

	Dependent variable:	
	n_cit_tot_log	
	OLS	*robust linear*
	(1)	(2)
n_authors_log	0.442***	0.407***
	(0.028)	(0.028)
n_references_log	0.323***	0.328***
	(0.033)	(0.032)
year	−0.451***	−0.460***
	(0.020)	(0.019)
month	−0.038***	−0.040***
	(0.005)	(0.005)
C(pre-print_match)	0.246***	0.242***
	(0.037)	(0.036)
C(software_used)	0.022	0.016
	(0.040)	(0.039)
C(software_created)	−0.084	−0.055
	(0.065)	(0.064)
C(software_shared)	0.175**	0.155**
	(0.068)	(0.067)
C(data_used)	−0.002	0.008
	(0.155)	(0.152)
C(data_created)	0.061*	0.053
	(0.033)	(0.033)

C(data_shared)	0.152^{***}	0.152^{***}
	(0.047)	(0.046)
C(bso_classification)Biology (fond.)	0.042	0.045
	(0.037)	(0.037)
C(bso_classification)Chemistry	−0.048	−0.057
	(0.184)	(0.180)
C(bso_classification)Computer and information sciences	0.161	0.125
	(0.110)	(0.107)
C(bso_classification)Earth, Ecology, Energy and applied biology	0.011	−0.026
	(0.068)	(0.067)
C(bso_classification)Engineering	−0.139	−0.096
	(0.111)	(0.109)
C(bso_classification)Humanities	0.036	0.016
	(0.080)	(0.078)
C(bso_classification)Mathematics	0.009	0.024
	(0.166)	(0.163)
C(bso_classification)Physical sciences, Astronomy	0.153	0.199
	(0.143)	(0.140)
C(bso_classification)Social sciences	0.201^{*}	0.174
	(0.121)	(0.119)
Constant	911.482^{***}	928.997^{***}
	(39.570)	(38.807)
Observations	2,812	2,812
R^2	0.271	
Adjusted R^2	0.266	
Residual Std. Error (df = 2791)	0.833	0.741
F Statistic	51.954^{***} (df = 20; 2791)	
Note:		*p<0.1; **p<0.05; ***p<0.01

References

[1] Aricia Bassinet et al. "Large-Scale Machine-Learning Analysis of Scientific PDF for Monitoring the Production and the Openness of Research Data and Software in France". 2023. (Visited on 05/08/2025).

[2] Iain Hrynaszkiewicz and Veronique Kiermer. "PLOS Open Science Indicators Principles and Definitions". In: (2022), 194697 Bytes. DOI: 10.6084/M9.FIGSHARE.21640889.V1. (Visited on 05/08/2025).

[3] UNESCO. *UNESCO Recommendation on Open Science*. Tech. rep. UNESCO, 2021. DOI: 10.54677/MNMH8546. (Visited on 05/08/2025).

[4] Open Science Monitoring Initiative. *Principles of Open Science Monitoring*. Tech. rep. Ministère de l'enseignement supérieur et de la recherche, 2024. DOI: 10.52949/49. (Visited on 05/08/2025).

[5] Ismael Rafols, Ingeborg Meijer, and Jordi Molas-Gallart. "Monitoring Open Science as Transformative Change: Towards a Systemic Framework". In: *F1000Research* 13 (Apr. 2024), p. 320. ISSN: 2046-1402. DOI: 10.12688/f1000research.148290.1. (Visited on 05/08/2025).

[6] S. Apartis et al. *Open Science Impact Indicator Handbook*. Jan. 2025. DOI: 10.5281/ZENODO.14538442. (Visited on 05/08/2025).

[7] Nicki Lisa Cole et al. "The Societal Impact of Open Science: A Scoping Review". In: *Royal Society Open Science* 11.6 (June 2024), p. 240286. ISSN: 2054-5703. DOI: 10.1098/rsos.240286. (Visited on 05/08/2025).

[8] Thomas Klebel et al. "The Academic Impact of Open Science: A Scoping Review". In: *Royal Society Open Science* 12.3 (Mar. 2025), p. 241248. ISSN: 2054-5703. DOI: 10.1098/rsos.241248. (Visited on 05/08/2025).

[9] Lena Tsipouri et al. *The Economic Impact of Open Science: A Scoping Review*. Feb. 2025. DOI: 10.31222/osf.io/kqse5_v1. (Visited on 05/11/2025).

[10] Giovanni Colavizza et al. "An Analysis of the Effects of Sharing Research Data, Code, and Preprints on Citations". In: *PLOS ONE* 19.10 (Oct. 2024). Ed. by Yongli Tang, e0311493. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0311493. (Visited on 05/08/2025).

[11] Public Library Of Science. *PLOS Open Science Indicators*. 2023. DOI: 10.6084/M9.FIGSHARE.21687686.V5. (Visited on 05/08/2025).

[12] Giovanni Colavizza et al. "The Citation Advantage of Linking Publications to Research Data". In: *PLOS ONE* 15.4 (Apr. 2020). Ed. by Jelte M. Wicherts, e0230416. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0230416. (Visited on 12/15/2023).

[13] Dominika Tkaczyk. *Crossref Relationships between Preprints and Journal Articles*. Nov. 2023. DOI: 10.5281/ZENODO.10144857. (Visited on 06/09/2025).

[14] Mariia Levchenko et al. "Enabling Preprint Discovery, Evaluation, and Analysis with Europe PMC". In: *PLOS ONE* 19.9 (Sept. 2024). Ed. by Florian Naudet, e0303005. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0303005. (Visited on 05/12/2025).

[15] Eric Jeangirard. "Monitoring Open Access at a National Level: French Case Study". In: *ELPUB 2019 23d International Conference on Electronic Publishing*. OpenEdition Press, June 2019. DOI: 10.4000/proceedings.elpub.2019.20. (Visited on 05/12/2025).

[16] Allison Langham-Putrow, Caitlin Bakker, and Amy Riegelman. "Is the Open Access Citation Advantage Real? A Systematic Review of the Citation of Open Access and Subscription-Based Articles". In: *PLOS ONE* 16.6 (June 2021). Ed. by Sergi Lozano, e0253129. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0253129. (Visited on 06/09/2025).

[17] Giovanni Colavizza, Lauren Cadwallader, and Iain Hrynaszkiewicz. *A Study on the Citation Impact of Open Science Indicators in the French Open Science Monitor*. 2025. DOI: 10.6084/M9.FIGSHARE.27822663.V2. (Visited on 06/10/2025).

[18] Patrick Vandewalle. "Code Sharing Is Associated with Research Impact in Image Processing". In: *Computing in Science & Engineering* 14.4 (July 2012), pp. 42–47. ISSN: 1521-9615. DOI: 10.1109/MCSE.2012.63. (Visited on 05/12/2025).

[19] Juan Pablo Alperin et al. *An Analysis of the Suitability of OpenAlex for Bibliometric Analyses*. 2024. DOI: 10.48550/ARXIV.2404.17663. (Visited on 05/12/2025).