Modelling the effect of funding selectivity on the uptake of data sharing in the academic community

Thomas Klebel[*], Federico Bianchi**, Tony Ross-Hellauer***, Flaminio Squazzoni**

[*]*tklebel@know-center.at*
0000-0002-7331-4751
Open and Reproducible Research Group, Know-Center GmbH, Austria

[**]*federico.bianchi1@unimi.it; flaminio.squazzoni@unimi.it*
0000-0002-7473-1928; 0000-0002-6503-6077
Department of Social and Political Sciences, University of Milan, Italy

****ross-hellauer@tugraz.at*
0000-0003-4470-7027
Institute of Interactive Systems and Data Science, Graz University of Technology, Austria

While the collective benefits of data sharing for science are clear, sharing data is not yet common practice in many research areas. Furthermore, there is scant knowledge on contexts and consequences of incentivising data sharing by funding agencies. Here, we built an abstract agent-based model to investigate the potential effect of funding selectivity and incentives for data sharing on the uptake of data sharing by academic teams which adapt strategically to resources. Our results suggest that more competitive funding schemes lead to higher rates of data sharing in the short run but lower uptake of data sharing in the long run than less selective funding. Attempts to reform systems of reward and recognition to foster Open Science practices should carefully consider the actual impact of measures and their potential long-term side effects.

1. Introduction

Sharing research data is considered beneficial to the community. It stimulates research reproducibility (Munafò et al., 2017), mitigates questionable research practices (Gopalakrishna et al., 2022), and can help to accelerate collective effort in response to unforeseen global events, such as during the COVID crisis (Tse et al., 2020). Articles which made their data available have been found to accrue more citations (Piwowar & Vision, 2013), and data sharing might lead to lower costs for accessing and using scientific data in the economy (Fell, 2019). Yet, data sharing is still not very common in many research areas (Serghiou et al., 2021).

Funding agencies have started to increasingly mandate or incentivise data sharing for grant applications and research output. Tedersoo et al. (2021) recently called for clear incentives for data sharing by providing actual benefits for researchers sharing data in promotion or grant funding decisions. However, there is still little understanding of the interplay between funding incentives, funding selectivity, and various further contextual factors on the diffusion of data sharing practices. Given that empirical data are unavailable to study these dynamics, we built an abstract, agent-based model that examines data sharing and grant seeking among a population of research teams. We modelled an environment where a fictious funding agency would incentivise data sharing for grant decisions and simulated aggregate consequences in terms of rate of data sharing. Here, we report intermediate results by focusing on the effect of a funding agency's level of selectivity on the uptake of data sharing among research teams which allocate their resources strategically.

1.1. Background

Policies for data sharing exist at multiple levels within the scientific system. Funding agencies increasingly encourage and sometimes mandate the sharing of primary research data (Gomes et al., 2022; Houtkoop et al., 2018). The European Commission is promoting FAIR (Findable, Accessible, Interoperable, Reusable) and Open Data in Horizon Europe, with explicit calls for research data management in compliance with the FAIR principles and to ensure access to research data following the principle of "as open as possible and as closed as necessary"[1]. At the level of journals and publishers, policies for data sharing are also becoming more common (Gomes et al., 2022; Vasilevsky et al., 2017).

Despite the increase in policies promoting data sharing, actual rates of data sharing remain low. Serghiou et al. (2021) investigated various aspects of transparency related to scientific publishing within the biomedical literature, including data sharing. Across 2.75 million Open Access articles from PubMedCentral, they found an increase in the rate of articles sharing data, with an estimated 15% of articles published in 2020 sharing research data. Similar rates of data sharing have been reported by Hamilton et al. (2022), who further found compliance with key FAIR principles to be extremely low.

However, low compliance with FAIR principles undermines the reusability of shared data. While journal policies mandating data sharing can increase rates of data sharing and enhance reusability (Hardwicke et al., 2018), offering authors to state that data were "available upon request" is often insufficient to ensure actual access to data (Tedersoo et al., 2021). In a study on all articles from Nature and Science between 2000 and 2019, Tedersoo et al. (2021) recommend data sharing to be associated with "real benefits such as recognition, or bonus points in grant and job applications".

Across countries and funding bodies, research funding is increasingly granted in the form of larger grants to fewer researchers, under the umbrella of "excellence" (Aagaard et al., 2020; Bloch & Sørensen, 2015). A systematic review of benefits and drawbacks of larger versus smaller grant sizes by Aagaard et al. (2020) found greater support for arguments in favour of smaller grants, related to aspects of efficiency, epistemology, and organisational aspects more broadly. Given that too small grants have been found to be inefficient as well, Aagaard and colleagues argue that the right "balance between concentration and dispersal" would depend on characteristics of scientific fields and national funding systems. In their model of the effect of Open Science interventions and differing funding schemes on rates of reproducibility, Smaldino et al. (2019) report that smaller grants are more effective at bringing about desired research practices. Since researchers depend on grants to survive academic competition, they would adhere more to the guidelines of funding agencies for small but repeated grants, whereas larger grants would lead to cumulative advantages (early success leading to later success), which could at least partly offset desired effects on research practices (Smaldino et al., 2019).

2. Method

To understand dynamics between incentives and selection standards of funding agencies, we built an abstract model that encapsulates the essential elements of grant allocation systems and scientific activities, inspired by previous models on the spread of poor methods (Higginson & Munafò, 2016; Smaldino & McElreath, 2016), and peer review dynamics (Bianchi et al., 2018). Our model consists of $n = 100$ academic teams that perform research and need grants. Teams are equipped with resources, which represent forms of capital such as available funds and previous publication success. Teams are initialised with a uniform distribution of resources, from low to high. Each round, teams receive a base-rate of resources, akin to the funding a PI

[1] https://research-and-innovation.ec.europa.eu/strategy/strategy-2020-2024/our-digital-future/open-science_en

has for their own position. In addition, teams seek to acquire research funding from one funding agency.

2.1. Funding allocation

To acquire funding, teams produce proposals by using resources. The quality of proposals depends on their resources, but not linearly. This mimics the fact that great proposals can be written by resource-poor teams, but on average more resources (e.g., more funds, better publication track record, better skill) lead to better proposals. The strength of each teams' proposal is drawn from a random normal distribution as follows:

$$\begin{aligned}
&normal(\mu, \sigma), \text{ where} \\
&mu = (1 - \text{sharing-incentive}) * \text{normalised-resources} + \text{sharing-incentive} \\
&\qquad\qquad * \text{sharing-effort} \\
&\sigma = 0.15
\end{aligned}$$

The model has a sharing-incentive parameter set to 0.4 that reflects an exogenous rule which promotes data sharing among funded research teams. The normal distribution and its parameter sigma ensure that there is no perfect path dependency to avoid that some teams would always be funded while others would not. Team resources are normalised to a [0, 1] scale. Similarly, a sharing-effort parameter follows the inverse logit scale [0, 1]. Furthermore, submitting proposals is assumed to be a costly activity. We assume that teams lose 5% of their resources each round to prepare proposals.

Based on the draws for their proposal strength, only top teams are eventually funded. The size of the total funding pool is twice as large as the base rate funding. Here, we vary the level of competition of the funding system, from a scenario where 10% of teams receiving large grants to a scenario where 60% of teams receive much smaller grants to mimic diverse funding schemes.

2.2. Data sharing

Teams are required to decide whether to share their research data. Teams are conceived as rational agents that face a competitive academic landscape. The probability of sharing data is determined in each round as follows:

$$\begin{aligned}
&bernoulli(p), \text{ with} \\
&p = \frac{1}{1 + \exp\left(-\text{sharing-effort}\right)}
\end{aligned}$$

Initial sharing-effort is sampled from a uniform distribution, with case (a) where there is a very low sharing effort, and case (b) where there is a of a uniform distribution across the scale of all potential efforts, thus mimicking a situation where there is an equal proportion of teams with low, moderate, and high sharing efforts. To adapt their sharing behaviour, teams compare resources between the current round (t_1) with the previous round (t_0).

Teams increase their sharing effort if (a) they shared data at t_0 and resources at t_1 are higher, or (b) if they did not share data at t_0 and resources are equal or lower at t_1. Otherwise, teams decrease their sharing effort.

Data sharing is understood as a costly activity. All teams have their resources reduced by up to 10% of their baseline funds. The value ϑ subtracted from teams' resources is calculated as follows:

$$\vartheta = 0.1 * \text{baseline-funding} * \text{inverse-sharing-effort}$$

where inverse-sharing-effort is the sharing-effort following a logit scale [0, 1]. This means if a team is investing a lot of effort into sharing data, their resources are reduced by 10% of their baseline funds. If a team does not invest any resources at all, no resources are subtracted.

Each simulated condition was run 100 times, and results were analysed in *R*. Table 1 shows the initial model parameters.

Table 1: Initial model parameters

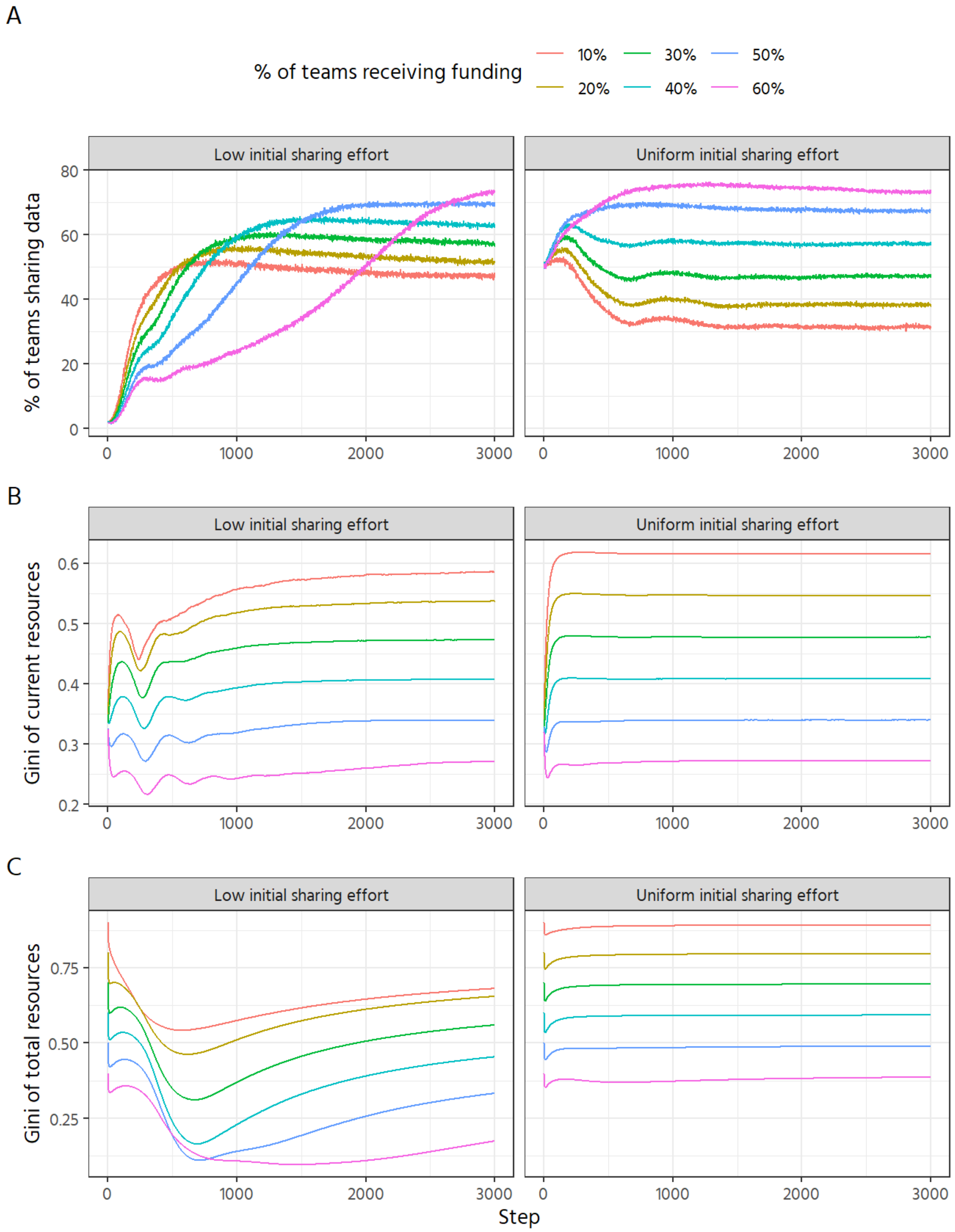
Parameter	**Initial value**
Number of teams	100
Sharing-incentive	0.4
Grant application penalty	0.05
Initial resource distribution	uniform
σ	0.15
Rate of third-party funding vs. base funding	2
Utility-change	0.03
Maximum of initial effort	Low initial sharing effort: -4 Uniform initial sharing effort: 4

3. Results

We considered three outcomes: (a) the percentage of teams currently sharing data; (b) the Gini coefficient of the current resource distribution; and (c) the Gini coefficient of overall resources. Calculating the Gini coefficients allowed us to gauge the general emerging system dynamics in terms of funded teams, and possible path dependency in funding decisions. Our preliminary analysis here will be backed up by more specific analysis at the level of individual research teams that will be presented at the conference.

We varied two input variables: (i) the share of teams receiving fundings, and (ii) the initial distribution of the sharing effort. Figure 1A shows that even without any exogenous variations of the incentives for data sharing from the fictious funding agency, sharing rate still greatly varied, depending on (i) the share of teams receiving funds, and (ii) the initial distribution of teams' sharing effort.

Figure 1: Effect of funding selectivity on rate of data sharing and resource distribution



When the initial sharing effort was uniform among teams, we assumed that the rate of data sharing started at 50%. When competitive selection for funds was high (only 10% of funded teams), data sharing declined to reach a stable equilibrium with only about 30% of teams sharing data. This would suggest that introducing sharing incentives in an academic environment where sharing is already quite common, but teams compete intensively for scarce funding opportunities, could lead to a *decrease* of data sharing. In less selective funding regimes, data sharing reached systematically higher levels, with up to 75% of teams sharing data. This is likely driven by the exposure of teams towards the funding agency: if more teams are funded, they are increasingly selected upon their data sharing practices, which in turn leads teams to increase their sharing effort. Inequalities in the resource distribution reached equilibria very quickly, reflecting the overall selective pressures set up by the fictious funding agency. When there is stronger competition for scarce funds, resources are distributed less equally at the system level.

When considering low initial sharing efforts, our model generated interesting and more realistic dynamics about the uptake of data sharing. The uptake of data sharing was fastest under the most competitive funding with large grants (only 10% of teams receiving funds each round), whereas it was consecutively slower for less competitive schemes (Figure 1A, left panel). However, the uptake of data sharing tapered off quickly in contexts of stronger competition for funds and reached lower levels of overall sharing in the long run compared to a scenario with smaller but less competitive grants.

In case of low initial sharing effort, dynamics related to the equity of resources as measured by Gini coefficients were markedly different. Although the general pattern of lower inequality with smaller grants was confirmed (Figure 1B left panel; Figure 1C left panel), its dynamics showed interesting outcomes, especially in the early stages of the simulation. More specifically, the Gini of total resources (Figure 1C) dropped substantially to never reach the levels of its counterpart of uniform sharing effort. This would suggest a higher turnover in terms of funded teams, and thus lower path-dependency. Therefore, this suggests that imposing policies to select teams partly based on their effort to share data could create multiple pathways towards success, where some teams would opt to share data while others would not. Our interpretation of this mechanism is only preliminary and will be backed up by a further analysis of individual level data.

4. Discussion

Our analysis tried to consider important dynamics between potential measures implemented by funding agencies to incentivise Open Science practices by academic teams, such as data sharing, an academic context where there is an existing uptake of these practices, and the selective pressure imposed by funding agencies upon funding competition. We found that more competitive funding schemes with larger grants lead to quicker uptake of data sharing but lower sharing in the long run, in particular when sharing is not common. In contrast, introducing incentives for data sharing in environments where data sharing is already common could decrease data sharing rates, if fund allocation is reasonably selective.

Our conclusions add evidence to the larger discourse on benefits and drawbacks of selectivity of research funding (Aagaard et al., 2020) and align with the findings by Smaldino et al. (2019) on the effect of grant size on changing academic practices. Results suggest that smaller grant sizes would be in the long run more effective in diffusing good practices endorsed by funding agencies. Highly selective funding schemes might lead to quicker uptake, but data sharing would need further support or incentives to stabilise in the long run. Attempts to reform reward and recognition in terms of Open Science practices, as recommended by Tedersoo et al. (2021), should therefore be considered carefully both in terms of actual impact and their potential side effects.

Of course, our results have various caveats. First, our results are only preliminary and require further analysis to be corroborated. Second, our results rely on a highly stylised simulation with certain necessarily simplified assumptions. For instance, our current analysis (a) does not include any networks between teams. In actual research fields, research teams are embedded in networks of collaboration and competition, which convey information about research and data sharing practices used to estimate others' behaviour. Network embeddedness is key to draw inferences on the most successful strategies to adapt to the environment and reduce uncertainty. In addition (b), teams can observe funding requirements *a priori*, thus potentially adapting their strategies to ensure their own success in the long run. In our model, teams "learn" funding requirements *post-hoc* based on their own success. Relatedly (c), research teams usually can obtain funds from multiple sources and are not bound to either their baseline funds or a single funding agency. Furthermore (d), our model is not calibrated on any empirical data that would help us to estimate Open Science policies, team behaviour and potential network effects. Finally (e), we do not consider scaling effects regarding a team's ability to share data. One might assume that research teams with more resources could more easily divert some of their own resources towards data sharing than smaller and/or less resourced teams. Our model currently does not consider such heterogeneous effects.

Future iterations of the analysis will incorporate different network topologies, representing exemplary research fields to calibrate context-specific factors. We will analyse various settings for funding agencies' incentives in more detail, also tracking individual teams' success trajectories to ground our interpretation on more micro-level evidence. We aim to publish these analyses as an expanded preprint and present them at the conference, substantiating the preliminary conclusions presented in this paper.

5. Conclusions

By modelling the effect of funding selectivity on the uptake of data sharing, we highlighted certain important contextual factors that could inspire current reform movements aimed at improving scientific practice through Open Science practices. Given highly selective funding, funder incentives might need to be complemented by other measures to achieve widespread adoption of data sharing.

Open science practices

The development of the model has benefited from earlier models being publicly available. Although our work is not complete and we plan to update the model and analysis, we share model code, simulation data and analysis code to facilitate peer-review and to foster transparency and reproducibility. All our materials are available at (Klebel et al., 2023).

Acknowledgments

We thank Jana Lasser for valuable feedback on the model and its implementation.

Author contributions

Thomas Klebel: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Software, Validation, Visualization, Writing - original draft, and Writing - review & editing.
Federico Bianchi: Conceptualization, Methodology, Software, Validation, and Writing - review & editing.
Tony Ross-Hellauer: Conceptualization, Funding acquisition, Methodology, Supervision, and Writing - review & editing.

Flaminio Squazzoni: Conceptualization, Methodology, Supervision, and Writing - review & editing.

Competing interests

We declare we have no competing interests.

Funding information

This work was supported by the project ON-MERRIT, funded by the European Commission under the Horizon 2020 programme (grant no. 824612). The Know-Center is funded within COMET—Competence Centers for Excellent Technologies—under the auspices of the Austrian Federal Ministry of Transport, Innovation and Technology, the Austrian Federal Ministry of Economy, Family and Youth and by the State of Styria. COMET is managed by the Austrian Research Promotion Agency FFG. Funded by the European Union. Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Commission. Neither the EU, nor the EC can be held responsible for them. FS was supported by a grant from the University of Milan (PSR2015-17 Transition Grant). FB was supported by a PRIN-MUR (Progetti di Rilevante Interesse Nazionale – Italian Ministry of University and Research) grant (Grant Number: 20178TRM3F001 "14All"). The funders had no role in the design of this study.

References

Aagaard, K., Kladakis, A., & Nielsen, M. W. (2020). Concentration or dispersal of research funding? *Quantitative Science Studies*, *1*(1), 117–149. https://doi.org/10.1162/qss_a_00002

Bianchi, F., Grimaldo, F., Bravo, G., & Squazzoni, F. (2018). The peer review game: An agent-based model of scientists facing resource constraints and institutional pressures. *Scientometrics*, *116*(3), 1401–1420. https://doi.org/10.1007/s11192-018-2825-4

Bloch, C., & Sørensen, M. P. (2015). The size of research funding: Trends and implications. *Science and Public Policy*, *42*(1), 30–43. https://doi.org/10.1093/scipol/scu019

Fell, M. J. (2019). The Economic Impacts of Open Science: A Rapid Evidence Assessment. *Publications*, *7*(3), 46. https://doi.org/10.3390/publications7030046

Gomes, D. G. E., Pottier, P., Crystal-Ornelas, R., Hudgins, E. J., Foroughirad, V., Sánchez-Reyes, L. L., Turba, R., Martinez, P. A., Moreau, D., Bertram, M. G., Smout, C. A., & Gaynor, K. M. (2022). Why don't we share data and code? Perceived barriers and benefits to public archiving practices. *Proceedings of the Royal Society B: Biological Sciences*, *289*(1987), 20221113. https://doi.org/10.1098/rspb.2022.1113

Gopalakrishna, G., Riet, G. ter, Vink, G., Stoop, I., Wicherts, J. M., & Bouter, L. M. (2022). Prevalence of questionable research practices, research misconduct and their potential explanatory factors: A survey among academic researchers in The Netherlands. *PLOS ONE*, *17*(2), e0263023. https://doi.org/10.1371/journal.pone.0263023

Hamilton, D. G., Page, M. J., Finch, S., Everitt, S., & Fidler, F. (2022). How often do cancer researchers make their data and code available and what factors are associated with sharing? *BMC Medicine*, *20*(1), 438. https://doi.org/10.1186/s12916-022-02644-2

Hardwicke, T. E., Mathur, M. B., MacDonald, K., Nilsonne, G., Banks, G. C., Kidwell, M. C., Hofelich Mohr, A., Clayton, E., Yoon, E. J., Henry Tessler, M., Lenne, R. L., Altman, S., Long, B., & Frank, M. C. (2018). Data availability, reusability, and analytic reproducibility: Evaluating the impact of a mandatory open data policy at the journal Cognition. *Royal Society Open Science*, *5*(8), 180448. https://doi.org/10.1098/rsos.180448

Higginson, A. D., & Munafò, M. R. (2016). Current Incentives for Scientists Lead to Underpowered Studies with Erroneous Conclusions. *PLOS Biology*, *14*(11), e2000995. https://doi.org/10.1371/journal.pbio.2000995

Houtkoop, B. L., Chambers, C., Macleod, M., Bishop, D. V. M., Nichols, T. E., & Wagenmakers, E.-J. (2018). Data Sharing in Psychology: A Survey on Barriers and Preconditions. *Advances in Methods and Practices in Psychological Science*, *1*(1), 70–85. https://doi.org/10.1177/2515245917751886

Klebel, T., Bianchi, F., Ross-Hellauer, T., & Squazzoni, F. (2023). *Code and data for 'Modelling the effect of funding selectivity on the uptake of data sharing in the academic community'*. Zenodo. https://doi.org/10.5281/zenodo.7851818

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., Simonsohn, U., Wagenmakers, E.-J., Ware, J. J., & Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*(1), 0021. https://doi.org/10.1038/s41562-016-0021

Piwowar, H. A., & Vision, T. J. (2013). Data reuse and the open data citation advantage. *PeerJ*, *1*, e175. https://doi.org/10.7717/peerj.175

Serghiou, S., Contopoulos-Ioannidis, D. G., Boyack, K. W., Riedel, N., Wallach, J. D., & Ioannidis, J. P. A. (2021). Assessment of transparency indicators across the biomedical literature: How open is open? *PLOS Biology*, *19*(3), e3001107. https://doi.org/10.1371/journal.pbio.3001107

Smaldino, P. E., & McElreath, R. (2016). The natural selection of bad science. *Royal Society Open Science*, *3*(9), 160384. https://doi.org/10.1098/rsos.160384

Smaldino, P. E., Turner, M. A., & Contreras Kallens, P. A. (2019). Open science and modified funding lotteries can impede the natural selection of bad science. *Royal Society Open Science*, *6*(7), 190194. https://doi.org/10.1098/rsos.190194

Tedersoo, L., Küngas, R., Oras, E., Köster, K., Eenmaa, H., Leijen, Ä., Pedaste, M., Raju, M., Astapova, A., Lukner, H., Kogermann, K., & Sepp, T. (2021). Data sharing practices and data availability upon request differ across scientific disciplines. *Scientific Data*, *8*(1), 192. https://doi.org/10.1038/s41597-021-00981-0

Tse, E. G., Klug, D. M., & Todd, M. H. (2020). Open science approaches to COVID-19. *F1000Research*, *9*, 1043. https://doi.org/10.12688/f1000research.26084.1

Vasilevsky, N. A., Minnier, J., Haendel, M. A., & Champieux, R. E. (2017). Reproducible and reusable research: Are journal data sharing policies meeting the mark? *PeerJ*, *5*, e3208. https://doi.org/10.7717/peerj.3208