ILLUSTRATION BY THE PROJECT TWINS

TAKING THE PAIN OUT OF DATA SHARING

Despite agreeing to make raw data available, some authors fail to comply. The right strategies and platforms can ease the task. **By Matthew Hutson**

Journals and funding bodies increasingly require manuscript authors to share data on request or make the information publicly available. It's a big ask from a technical standpoint, but some straightforward strategies can simplify the process.

Scientific papers rarely include all the data used to justify the conclusions, even in the supplementary material. Authors might fear getting scooped, or that other researchers will use the raw data to make fresh discoveries, or they might wish to protect the privacy of study participants. Or, more probably, authors have neither the time nor the expertise to package the data for others to view and understand.

Such reticence costs the research community. Data transparency allows others to repeat analyses and catch mistakes or fraudulent claims. It allows for new findings through the reanalysis of existing data sets, and it increases trust in the scientific process. In August, the White House Office of Science and Technology Policy announced that, by 2025, scientific data from all new federally funded research must be made accessible to the US public. And when submitting papers, authors are increasingly required to provide raw data to editors, to place data online or to include data-sharing statements as to whether they will offer data on request. Unfortunately, such policies are not bulletproof, as the largest study of its kind starkly documents.

In May, Livia Puljak, who studies evidence-based medicine at the Catholic University of Croatia in Zagreb, and her colleagues published a study in which they looked at the roughly 300 journals published by BioMed Central, an open-access publisher that is part of Springer Nature, which also publishes *Nature*. The researchers identified 1,792 manuscripts published in January 2019 that declared their data were available "on request" or "on reasonable request"[1]. In early 2021, they e-mailed the corresponding authors, asking for access to the raw data. To allay concerns that the study could produce embarrassing findings, they noted that the analysis would be anonymized: "We will not disclose any details about author characteristics," they wrote.

Two hundred and fifty-four authors replied, of whom 123 shared their data. Among respondents who did not share data, the most common circumstances were that they asked for more information and then went silent

when provided with it (17%); they said they were not allowed to share the data (11%) or could not access the data (9%); or they offered no explanation (8%). The study, published in the *Journal of Clinical Epidemiology*, does not have publicly available raw data because the authors did not want to publicly shame other authors.

All journals in the study required the authors to state whether they would share their data. But because sharing wasn't a condition of publication, it's unclear why the authors who did not intend to share their data didn't simply say so. "Maybe they were giving socially acceptable answers," Puljak says. "Probably, people don't really think about what will happen when somebody actually asks for data."

Tom Jefferson, an epidemiologist at the University of Oxford, UK, says authors should face consequences for making false data-availability statements. "The editors should take action, whether it's a correction or retraction," he says, adding that the excuse of no longer having the data to hand is like saying "the cat ate my filing cabinet". But David Mellor, director of policy at the Center for Open Science (COS) in Charlottesville, Virginia, is not a fan of retraction. "It's kind of a blunt instrument," he says. Referring to the study's findings, he notes, "there's a possibility that the e-mail was simply not seen."

Valentin Danchev, a computational social scientist at Queen Mary University of London, calls the study a useful step towards understanding the actual state of data sharing. But, he adds, "we need more of those studies so that we can generalize across different areas and different survey designs".

Last year, Danchev co-authored a study[2] of 487 clinical trials that were published in *JAMA*, *The Lancet* or *The New England Journal of Medicine*. The authors of 89 of these articles said they'd stored data sets in online repositories, but Danchev's team could find only 17 in the designated locations.

In 2020, Tsuyoshi Miyakawa, a behavioural neuroscientist at Fujita Health University in Toyoake, Japan, and editor-in-chief of *Molecular Brain*, wrote in an editorial[3] that, since 2017, he had asked the authors of 41 papers for raw data before publication, because he'd felt the submitted data were "too beautiful to be true". The authors of 21 of those papers withdrew their submissions, and he rejected 19 of the rest on grounds of insufficient data. The experience made Miyakawa something of a sceptic: in the editorial, he proposes that editors stop assuming that researchers are honest.

Data definitions

Reforms might need to come from the top, researchers suggest. Puljak and her co-authors say they wish the practice of requiring authors to submit raw data before publication was more widespread. They are not alone. Several researchers contacted by *Nature* said that journals bear some of the blame for prioritizing original research and subscription fees over the policing of data sharing. When asked if publishers bear any responsibility for ensuring that authors follow through on their data-sharing statements, Chris Graf, research-integrity director at Springer Nature, said: "It is the author's or their institution's responsibility to honour author statements about data sharing." (*Nature*'s journalism is independent of its publisher.)

'Data sharing' means that data can be obtained from the study authors on request; the related but distinct concept of 'open data' means that the data are broadly accessible through online repositories and related resources. Puljak's study, Mellor says, is "an indication that 'available upon request' doesn't cut it". 'Open data' performs better: a study[4], published in August, of papers that had appeared in *PLoS ONE,* found that 88% of the data-availability statements containing URLs or DOI codes contained sufficient information to retrieve the data.

"People don't really think about what will happen when somebody actually asks for data."

Multidisciplinary repositories such as Figshare (part of Digital Science, which, along with Springer Nature, is part of the Holtzbrinck Publishing Group), Zenodo (operated by CERN, Europe's particle-physics laboratory near Geneva, Switzerland) and OSF.io (operated by the COS) are popular options for data deposition. But "any repository that's specifically designed for the type of data you are generating is probably the best", Mellor says. Such repositories, such as HEPdata for high-energy physics and OpenNeuro for neuroimaging, often format data to community standards and make them discoverable by researchers in those fields. Some also have protocols for protecting sensitive data such as medical records.

The Inter-university Consortium for Political and Social Research (ICPSR), an organization that supports open social science, offers extensive, professional data curation, says Amy Pienta, a researcher focusing on the demographics of ageing at the University of Michigan, Ann Arbor, where the ICPSR is based. ICPSR curators check for missing data, review data quality and create a dictionary of data labels. To preserve participant privacy, they might remove identifying information or restrict access to authorized users.

Pienta recommends that researchers who curate their own data follow these kinds of steps as well. "The reusability of data comes from creating metadata in an organized-enough way that somebody can understand the study without your looking over their shoulder," she says. Think carefully, even about the file format, she adds. Some journals allow supplementary files only in the form of PDFs, for instance. "That is a nightmare," Puljak says, because the format can make it difficult to extract data for subsequent analysis.

European Union-funded projects such as OpenAIRE, FOSTER Plus and Orion provide training materials on open science, including workshops, guidebooks and online courses.

Thinking ahead

According to Mellor, if researchers wish to enhance data availability, they need to make it less of an afterthought. "It's very tempting to address data sharing as a last step in the process, one that's not too important," he says. "And that motivates a lot of our efforts to focus on the beginning of the research process."

Study preregistration, in which authors share their experimental and analytical protocols before starting their analyses, to discourage selective publication of positive results and poor statistical practices such as '*P*-hacking', is intimately connected with open data and methods, Mellor says. "At the beginning of a study, asserting precisely how data will be collected and preserved, and what hypotheses are going to be tested – that really sets one up for success. Then it can just be a matter of filling in the bits as the data are collected."

The COS maintains a study registry at OSF.io. The organization has also created badges to indicate preregistration, as well as open data and open materials, which participating journals can place on papers. More than 120 journals currently display the badges, Mellor says, thereby normalizing open science by showing that "it's not as weird or as out-there as we often think".

Forming a data-sharing plan ahead of time also "makes your own science better", Pienta says. And sharing data and open data can lead to increased citations of your own papers or even to co-authorship of papers if other researchers use the data. Data sets stored in repositories can receive their own DOIs, turning them into stand-alone publications that might please grant and tenure committees.

Beyond those advantages, there's the gratification that comes from making your research widely accessible, particularly during a pandemic. "If somebody is collecting data on the new disease and sharing it," Puljak says, "it could help the whole world."

Matthew Hutson is a freelance science writer based in New York City.

1. Gabelica, M., Bojčić, R. & Puljak, L. *J. Clin. Epidemiol.* **150**, 33–41 (2022).
2. Danchev, V., Min, Y., Borghi, J., Baiocchi, M. & Ioannidis, J. P. A. *JAMA Netw. Open* **4**, e2033972 (2021).
3. Miyakawa, T. *Mol. Brain* **13**, 24 (2020).
4. Federer, L. M. *PLoS ONE* **17**, e0272845 (2022).

Correction

This Technology feature incorrectly stated that Figshare is owned by Springer Nature. In fact, it is part of Digital Science, a firm operated by the Holtzbrinck Publishing Group, which has a share in Springer Nature.

Corrected 4 October 2022