READY, SET, SHARE!

As funders roll out new requirements for making data freely available, researchers weigh costs and benefits

By **Jocelyn Kaiser** *and* **Jeffrey Brainard**

Physiologist Alejandro Caicedo of the University of Miami Miller School of Medicine is preparing a grant proposal to the U.S. National Institutes of Health (NIH). He is feeling unusually stressed because of a new requirement that takes effect this week. Along with his research idea, to study why islet cells in the pancreas stop making insulin in people with diabetes, he will be required to submit a plan for managing the data the project produces and sharing them in public repositories.

For his lab, that's a daunting task. Unlike neuroscience or genomics, Caicedo's field has no common platforms or standards for storing and sharing the kinds of data his lab generates, such as videos of pancreatic islet cells responding to a glucose stimulus. The "humongous" raw imaging files are currently stored in an on-campus database, notes Julia Panzer, a postdoctoral researcher in the lab. To protect patient privacy, the database is secured and not designed to provide access to outsiders. Sharing the data will mean uploading them somewhere else.

Caicedo supports the new NIH policy, acknowledging that "science will be so much more powerful" if data are freely shared. But he says his field isn't ready. And he's worried about the burden the new mandate will impose on his postdocs and graduate students. He can't afford to hire a data manager for his eight-person lab with his $600,000 in NIH grants, he says. "It's a very limited budget for a lot of people."

In the years ahead, many researchers will be struggling with similar issues. By 2025, new U.S. requirements for data sharing will extend beyond biomedical research to encompass researchers across all scientific disciplines who receive federal research funding. Some funders in the European Union and China have also enacted data-sharing requirements. The new U.S. moves are feeding hopes that a worldwide movement toward increased sharing is in the offing. Supporters think it could speed the pace and reliability of science.

Some scientists may only need to make a few adjustments to comply with the policies. That's because data sharing is already common in fields such as protein crystallography and astronomy. But in other fields the task could be weighty, because sharing is often an afterthought. For example, a study involving 7750 medical research papers found that just 9% of those published from 2015 to 2020 promised to make their data publicly available, and authors of just 3% actually shared, says lead author Daniel Hamilton of the University of Melbourne, who described the finding at the International Congress on Peer Review and Scientific Publication in September 2022. Even when authors promise to share their data, they often fail to follow through. Out of 21,000 journal articles that included data-sharing plans, a study published in *PLOS ONE* in 2020 found, fewer than 21% provided links to the repository storing the data.

Journals and funders, too, have a mixed record when it comes to supporting data sharing. Research presented at the September 2022 peer-review congress found only about half of the 110 largest public, corporate, and philanthropic funders of health research around the world recommend or require grantees to share data.

"Health research is the field where the ethical obligation to share data is the highest," says Aidan Tan, a clinician-researcher at the University of Sydney who led the study. "People volunteer in clinical trials and put themselves at risk to advance medical research and ultimately improve human health."

Across many fields of science, researchers' support for sharing data has increased during the past decade, surveys show. But given the potential cost and complexity, many are apprehensive about the NIH policy, and other requirements to follow. "How we get there is pretty messy right now," says Parker Antin, a developmental biologist and associate vice president for research at the University of Arizona. "I'm really not sure whether the total return will justify the cost. But I don't know of any other way to find out than trying to do it."

Science offers this guide as researchers prepare to plunge in.

ILLUSTRATION: DAVIDE BONAZZI/SALZMANART

Why share scientific data?

Scientists who share data can speed up and improve science and advance their own careers, advocates say.

Growing efforts in psychology, cancer research, and other fields to reproduce published studies, for example, depend on access to the underlying data. Such access can help resolve whether an apparent error in a paper resulted from an honest mistake—or faked data. In 2020, for example, a high-profile study claiming that COVID-19 patients given the antimalarial drug hydroxychloroquine were at increased risk of death fell apart after Surgisphere, the company that claimed to have provided the underlying data set, couldn't produce it. (Studies purporting to show benefits from the drug, touted by then-President Donald Trump, proved equally problematic.)

Sharing data could also help curtail duplicative efforts to collect them. That could save time and money for smaller labs in particular, says Crystal Rogers, a cell and developmental biologist at the University of California (UC), Davis. "Maybe this policy will even the playing field," she says. "It will democratize opportunities."

Existing data can help researchers generate hypotheses, design clinical trials, and teach. And by pooling smaller data sets, scientists can conduct meta-analyses that can produce robust or intriguing findings, says Maryann Martone, a neuroscientist at UC San Francisco. She points to a study that gathered raw data from an array of animal studies conducted in the 1990s on treatments for spinal cord injury. The results from the individual studies were inconsistent and never published. But a 2021 analysis of pooled data from 1125 animals produced a significant correlation: Animals with blood pressure levels within a certain window during spine surgery fared better, a finding that held up in a clinical study. "There's real gold in these small data sets, if you can put them together," Martone says.

For the researchers who share their data, one proven reward is increased citations to papers for which data are provided. Papers that provided a link to data gained 25% more citations on average than those that did not, according to a 2020 study of more than 50,000 articles in the PLOS and BMC journals.

Even as more funders expect grantees to provide data, a lack of professional rewards may be responsible for widespread noncompliance. Sharing typically doesn't count for much in tenure and promotion reviews, for example. Academic institutions should encourage departments to develop policies for providing such rewards, according to a 2021 joint report from the Association of American Universities and the Association of Public and Land-grant Universities.

It may be hard to overcome fears that researchers who share data won't get proper credit from others—or may even get scooped. "How do you make sure that somebody doesn't grab that data and publish it as their own in some minor journal?" worries cancer physician-scientist Jan Grimm of Memorial Sloan Kettering Cancer Center. Advocates for data sharing have called for publishers to discourage such behavior by requiring authors who use data generated by other scientists to name them as "data authors."

Scientists may come to see data sharing as a useful burden, like peer review, says Tim Vines, founder of a data search tool called DataSeer. "Peer review is very annoying, but many people say: 'It improves my manuscripts.' Researchers accept that. We need to bring [data] sharing to that level."

How are data-sharing policies changing?

Many U.S. funders already have sharing policies. NIH has been a leader of such efforts, rolling out a 1996 policy for its grantees in human genome sequencing and expanding it in 2003 to cover all large projects. Now, the agency is extending its rule to cover all of its research grants.

NIH's new policy "strongly encourages" researchers to deposit project data in repositories where other researchers have free access. The data should be "of sufficient quality to validate and replicate research findings," the policy says. Data should be deposited when a journal article about them is published or the grant ends, whichever comes first. And the policy extends to unpublished findings, including negative results.

"We really wanted to catalyze the research community through a more ubiquitous data-sharing policy," says Lyric Jorgenson, acting director of NIH's Office of Science Policy, who oversaw development of the policy.

NIH's push could test the kinds of changes that all federally funded researchers will need to make by December 2025, when a revised data-sharing policy announced in August 2022 by the White House Office of Science and Technology Policy (OSTP) takes effect. (The new policy also drew attention for requiring that journal articles be free to access when published.)

Like NIH's policy, OSTP's requires all data "underlying" peer-reviewed scholarly papers be made publicly available for free when papers are published (although it allows exceptions, including if taking that step is too costly). The policy is significantly more stringent than the current requirement at a key agency it covers, the National Science Foundation (NSF), which only requires data sharing within "a reasonable time period." NSF and other U.S. research-funding agencies are expected to propose details this year and next about how they will implement OSTP's policy.

How do I comply?

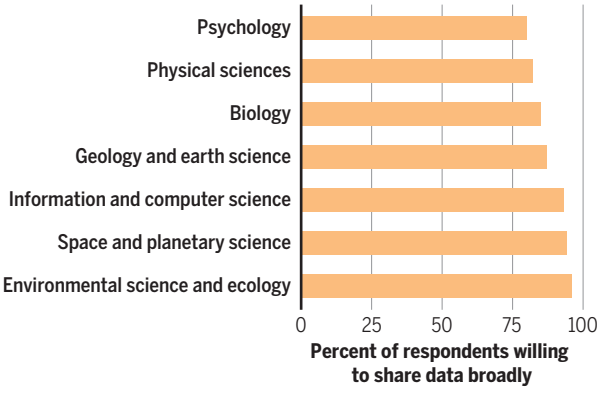
NIH's policy requires investigators submitting a grant proposal to include a two-page data-management plan listing the types of data they will produce, the software or tools needed to use the data, and the publicly accessible repositories where they will be stored. When submitting the data, researchers will need to include "metadata," or details of how the data were collected. And they may need to reformat data to fit a repository's standards. These steps are meant to make the data comply with international guidelines called the FAIR principles, which stands for "findable, accessible, interoperable, and reusable."

Supporters of sharing also call for something encouraged, but not required by NIH: choosing repositories that attach digital object identifiers (DOIs) to data sets that are different from those used to identify the associated papers. The DOIs—unique, permanent serial numbers—will make it easier for other researchers to find relevant data. (Authors and journals must also properly format a manuscript's references to associated data for them to be discoverable by search tools.) DOIs will also identify each data set as an independent scholarly contribution, enabling researchers to claim credit for generating and sharing the data.

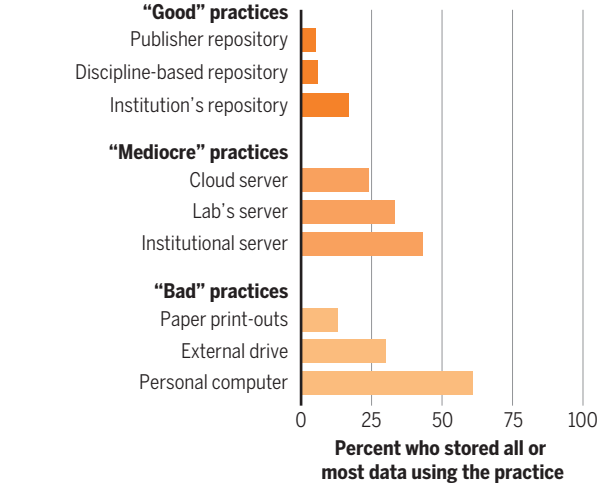
ILLUSTRATIONS: A. MASTIN/*SCIENCE*

Data sharing, by the numbers

Scientists express interest in sharing their data ...
Across fields, most scientists like the idea of sharing data, according to a 2017–18 survey of more than 2000 respondents from multiple countries. (Selected categories are shown.)



... but few do it right, if at all.
When researchers do share data, many don't use practices that promote long-term curation and sharing with other researchers. Analysts say big changes in research culture are needed, and more training and guidance could help. (Survey respondents could report more than one type of storage method.)



What would motivate researchers to share data?
Professional rewards rank highest, but researchers also see public benefits, according to a survey of more than 6000 respondents in 2022. (Selected motivations are shown, and respondents could choose more than one reason.)

Some researchers who recently began to share more data under existing NIH policies say the process can be very time-consuming. Relabeling, reformatting, and otherwise preparing all the underlying data collected by co-authors on a paper can take half a day, says Florian Krammer, a virologist at the Icahn School of Medicine at Mount Sinai—work he typically does on a weekend. His data manager needs another full day to upload the data to databases. "I think a lot of people don't realize how much work it is," he says.

Others point out that if researchers develop plans for data sharing at the start of a project, the costs may go down. "The time decreases the better managed your lab is, because things are documented from the get-go instead of at the very end," UC San Diego's Martone says. Stacey Schultz-Cherry, a virologist at St. Jude Children's Research Hospital, puts it this way: "We're all going to grumble, but in the long run it's really going to benefit science."

What about sensitive data that are difficult to share?

Biomedical research often involves human subjects who may not have agreed to having their data shared, even if they have been stripped of identifying information. NIH's policy allows exceptions for such data. But the agency expects that, when possible, consent forms for new studies will ask participants to agree to share their deidentified data.

Although some institutional ethics boards have opposed such broad consent, "I think there's an understanding that this is what the community is moving towards," says physician Ida Sim of UC San Francisco. She is a co-founder of Vivli, a repository that some institutions plan to use to share participant-level data from clinical trials.

Clinical trial researchers are known for keeping their data under wraps because they're concerned they won't be analyzed properly, or they're still writing papers. Many already ignore a 2016 NIH rule requiring that summary data be posted in the federal ClinicalTrials.gov database no later than 1 year after the trial's primary completion date. But the NIH data-sharing policy is already "being taken as a serious mandate" by those scientists, Sim says. "I am pleased with how much of a cultural change this has catalyzed."

How will policies be enforced?

Because biomedical disciplines create and use data differently, NIH says it chose to provide flexibility by not packing its policy with detailed requirements.

In particular, it does not specify how much data researchers must share from a given data set. Do they need to deposit an entire video of dividing cells or a molecular marker infiltrating a tumor, which could be gigabytes of data, or just the still images presented in papers? "Many of us do not fully understand at what level, from raw to fully processed and grouped, NIH expects data to be shared," says cardiovascular disease researcher Curt Sigmund of the

CREDITS: (ILLUSTRATIONS) A. MASTIN/*SCIENCE*; (GRAPHICS) K. FRANKLIN/*SCIENCE*; (DATA, TOP TO BOTTOM) C. TENOPIR *ET AL. PLOS ONE*, 15(3), E0229003(2020) (2); G. GOODEY *ET AL., THE STATE OF OPEN DATA 2022*; DIGITAL SCIENCE REPORT. HTTPS://DOI.ORG/10.6084/M9.FIGSHARE.21276984.V5

Medical College of Wisconsin. The answer, NIH's Jorgenson says, is that each discipline will need to work out the "granularity" required to reproduce a paper's findings.

In practice, NIH program managers will review an investigator's sharing plan when a grant proposal is submitted and check progress reports to be sure the plan is being followed. The agency could terminate a grant for noncompliance, although that rarely happens for violations of other NIH policies. But those who don't share data could be barred from receiving a new grant, Jorgenson says. Achieving the data-sharing policy's goals will likely be achieved "in stages and steps," she adds. "We did not want to set the bar so high that we create disincentives for anyone to participate."

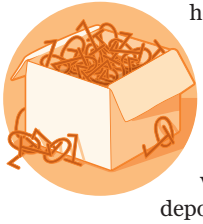
Many funders and journals have struggled to enforce their own sharing requirements. Confirming whether authors shared all data supporting an article can require a close, time-consuming examination, Vines says. Publishers receive no extra revenue for the added effort.

To avoid data sharing that is incomplete or poorly done, funders and institutions may need to not only threaten researchers with sticks, but also offer them carrots, in the form of technical support and training, says Dylan Ruediger, a project manager at Ithaka S+R, a higher education research and consulting organization. He managed an NSF-funded project that brought together interdisciplinary teams of researchers in fields as disparate as agronomy, nuclear imaging, and polar science to examine barriers to data sharing.

"Complying with mandates to deposit data is not the same thing as creating an ecosystem that's really well adapted to help researchers reuse data," Ruediger says. "That's a very different kind of challenge."

How and where do I store data?

Currently, many researchers store data primarily on their personal computers (see graphic, p. 324). Sharing data will mean shifting them to one of several possible homes: a repository at the researcher's institution; a discipline-based one, such as OpenNeuro, which holds brain-imaging data, or NIH's ImmPort, which stores immunology data; or a general repository, such as figshare or Zenodo. Many repositories will need improvements to make it easier to deposit, find, and retrieve data, experts say.



To help navigate this new terrain, some universities are beefing up staff who can help, such as IT specialists and librarians who specialize in data. "We're reprioritizing some of the things that we're doing in the library to accommodate these requests," says Vicki Coleman, dean of library services at North Carolina A&T State University, a historically Black research institution. She says the library will shift staffing away from its traditional reference desk—a trend underway at other universities as well.

These data experts often have clever ways to adapt commonly used information-management tools to the needs of specific research fields. Many universities, for example, now offer faculty members training in using Jupyter Notebooks, an open-source web application designed to make it easier to share data. The extra staffing and training should address a concern Ruediger found among participants in his project to encourage data sharing: "a sense that the challenges they were facing were unique and idiosyncratic to them."

What are the costs, and who will pay?

Scientists say it can be difficult to estimate the cost of cleaning and preparing data for use outside their team. For example, Krammer of Mount Sinai estimates data sharing will eat up at least 10% of his funding. Hiring a data manager might cost $100,000 per year, although not all labs will need one.

NIH says researchers applying for a grant can add costs for data managers, staff time to prepare data, and repository fees. But because NIH has a strict dollar limit for many grants, data-sharing costs may cut into the funds available for research. "If you're loading up your grant budget with data-sharing and management costs, is that going to take away from the funds for doing the science?" says David Kennedy, vice president of the Council on Governmental Relations, which represents major research universities.

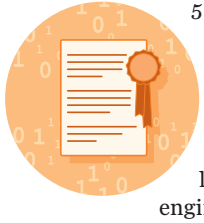
Universities, for their part, will have to pay for campus-wide services supporting data sharing, such as librarians and subscriptions to repositories. Institutions can bill these "indirect" costs to the overhead that funders provide with grants. But those reimbursements are capped. Although universities have long tapped their own revenues to help cover the indirect costs of research, some worry data sharing will become another "unfunded mandate" from the federal government.

The costs per institution will exceed $1 million a year, split between overhead and investigators' budgets, according to an initial analysis based on a survey of 34 Council on Governmental Relations members. That could be a special burden for smaller institutions, Kennedy says. That is "a huge concern," Jorgenson acknowledges. "We do not want to exacerbate inequities in the funding structure."

Another challenge to be solved: Even the largest repositories are still looking for sustainable business models. Discipline-specific ones are typically supported by grants for individual projects that don't assure funding after the grant ends. NIH's and OSTP's policies don't spell out for how long data must be stored and shared; Jorgenson says the agency "will be collecting lots of information" to inform a more specific policy on this.

Will broad data sharing be worth the effort?

Skeptics say the benefits are yet to be demonstrated. Krammer says funders should collect and analyze data about whether the new push is producing the intended effects. "There needs to be an evaluation after 2 years, 5 years, to look at what type of data is [re]used, and for what type of data it doesn't seem to make sense," he says.



Supporters of data sharing agree—and think the results will bear them out. "We need some real demonstrations of how this level of data sharing can drive the discovery engine," UC San Francisco's Sim says. "I don't think we're there yet. But it's kind of like everyone's hopped in the car, and we're starting the engine." ■

ILLUSTRATIONS: A. MASTIN/*SCIENCE*