https://doi.org/10.1038/s41592-023-01927-7

To share is to be a scientist

Wrangling big data is now part of being a biomedical scientist, and mandates on data sharing have entered the scene. Mandates can alter behavior, but data sharing also needs incentives and shifts in science culture.

By Vivien Marx

As data-spewing instruments spread across biomedical labs and multimodal approaches are embraced, data sharing must be powered up, too. Much has been achieved, say some researchers in genomics, proteomics, neuroscience and imaging, as do some big data producers, wranglers at repositories and shepherds of large-scale projects. Big biodata's next phase, they say, needs resources and shifts in science culture. Here are some views on how far things have come and what lies ahead.

With data as with pizza, it's considered good manners to share. Whereas pizza sharing is a private affair, data sharing is how good citizens of science give collaborators and strangers access to results generated mainly or entirely with public funds. As of January 2023, the US National Institutes of Health (NIH) mandates all who apply for funding must submit a Data Management and Sharing plan. The reaction has not been a chorus of hurrays.

Some tense moments occurred at this year's American Association for Cancer Research (AACR) annual meeting, when program officers from the National Cancer Institute's (NCI) Office of Data Sharing presented the new NIH data sharing policies and held a question-and-answer session.

Some investigators said the mandate had been rolled out too suddenly. They asked how they are to find the time, skill and funding to set up a data management and sharing plan. Program officers directed attendees to guidance pages and offered personal conversations. Scientists can send questions to program officers or, in the case of NCI, to NCIOfficeOfDataSharing@mail.NCI.gov. The first grant proposals with data management and sharing plans have been submitted, the NCI Office of Data Sharing said in an unattributed statement after the conference. The plans are still in study section and review, so it's unclear how much back-and-forth will unfold. Says Emily Boja from the NCI Office of Data Sharing, who presented at the AACR session, it's understandable that the mandate can seem overwhelming for labs not used to organizing data for sharing, "especially when resources are limited."

The culture shift we are striving for is to weave data management and sharing into the conduct of science, says Heather Basehore.

CREDITS: TOP: C. JOHNNY / GETTY IMAGES; RIGHT: NIH NCI OFFICE OF DATA SHARING

Before COVID-19, sequence databases held 30,000 coronavirus sequences, which sped the study of SARS-CoV-2, the virus that caused the pandemic, says Guy Cochrane.

Heather Basehore from the NCI's office of Data Sharing, who also presented, hopes researchers don't consider their data management and sharing plan as "just a box that needs to be checked off," or a task that can be left to the last minute. "Even though we understand that this feels like a burden, the culture shift that we are striving for is for data management and sharing to be woven into the conduct of science, so if you are thinking about this at the beginning as part of your research process, you will hopefully make adjustments to the way you design and conduct the experiments at the outset."

What scientists fear, says Jason Swedlow, a quantitative cell biologist at the University of Dundee who was present at the AACR session, is the 'unfunded mandate' or the 'do more with less' approach. But to his ears, says Swedlow, NCI is willing to fund some of the data-sharing activities. Although some labs make their own arrangements for data sharing, in general, in his view, universities and other academic institutions and non-profit research institutes are "not geared up for this, and their IT infrastructures and staff are not designed for these types of activities." He was glad to hear the NCI program officers say: 'talk to us'. "It's going to be a conversation," he says.

Swedlow sees new concepts underpinning the mandate: namely, that publicly funded research should be published and it's public property, not the property of the researcher who did the experiment. This is understood in genomics, structural biology and imaging, but the NIH is applying this to all research data, he says. "That's quite a step." Another new concept is that data are more than a figure in one's paper, he says. One scientist at the AACR event said, rather vociferously, that he had been publishing data throughout his career: namely, in the figures of his papers. He left the room after saying this. But, says Swedlow, "I think the trend is that a figure is not 'data'."

European data-sharing mandates are probably less structured than the NIH ones right now, says Mallory Freeberg, who coordinates the European Genome-phenome Archive (EGA), which is hosted at the European Molecular Biology Laboratory/European Bioinformatics Institute (EMBL-EBI). It holds personally identifiable genetic, phenotypic and clinical data. She sympathizes with those struggling with a data management and sharing plan. In the excitement of starting a new project, she says, it might seem tedious to first tend to data management. But waiting until a study is completed is not the best way to package data for sharing through archives.

It's noble

One of his favorite examples about the benefits of data sharing, says Guy Cochrane from the EBI, is the fact that before the COVID-19-pandemic, genomic sequence databases held 30,000 coronavirus sequences. These data sped up identification and characterization of SARS-CoV-2, the virus that caused the pandemic, says Cochrane, who heads the European Nucleotide Archive (ENA), a global repository for sequence data. The ENA holds DNA and RNA sequence data and mirrors GenBank in the United States and the DNA DataBank of Japan.

Data sharing is "near and dear to my heart," said Monica Bertagnolli, director of the NCI, in response to my question during a press conference at the AACR annual meeting. She has been nominated to direct the NIH. Said Bertagnolli, when patients donate their data to research, they want them to be used, in trust, because they understand the benefit the data can bring.

The NIH funds fundamental discoveries, research strategies and their application to advance human health. This goal, says John Quackenbush, computational biologist at the T.H. Chan Harvard School of Public Health, is not met "if nobody else can replicate my results," if results are somehow suspect or "if I make a mistake that somebody else can't find." Papers that share data are cited more often than papers that do not, and in papers presenting a software tool, when data are provided, more people will use the software tool and cite the paper, he says. With data sharing, "genomics has done generally better than most other fields," he says, which is in part due to data release practices during the Human Genome Project. This tradition has continued and infrastructure has been built to enable data access.

When people held back data, we said you've got to play by our rules, says Anna Barker.

Past efforts reveal the community-wide lasting benefits of such behavior. As former deputy director of the NCI and former deputy director for strategic initiatives, Anna Barker is familiar with less than cheering reactions from the research community about data-sharing mandates. She is now chief strategy officer at the Ellison Institute for Transformative Medicine in Los Angeles. Half of NCI-funded research is done by individual investigators who care about sharing data but now probably feel rather alone as they face this data management and sharing mandate, says Barker.

Big labs and institutes such as the Broad Institute of MIT and Harvard have long practiced data sharing and did so along with other investigators involved in The Cancer Genome Atlas project (TCGA), a program Barker co-initiated and shepherded with colleagues at the NIH National Human Genome Research Institute. TCGA's 2.5 petabytes of data stem from molecular characterizations of 20,000 cancer tissues from 33 cancer types and include healthy tissue samples from the same people. Data sharing was mandated and it worked, she says. When people held back data, "We said, 'No, if you want to play with us, you've got to play by our rules'."

When TCGA was launched in late 2005, genomics labs had achieved much, especially related to cancer genes. But if work was to continue by amassing cancer gene mutations one at a time, it would take decades to compile and characterize cancer genomes. At the same time, among cancer researchers, TCGA faced an "anti-big science" undercurrent, says Barker. The project shows how investigator-driven and large-scale projects reinforce one another. Big science enables

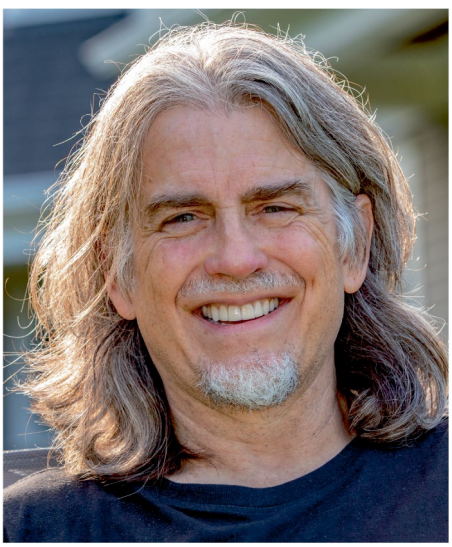
CREDITS: LEFT: EMBL-EBI; RIGHT: ELLISON INSTITUTE FOR TRANSFORMATIVE MEDICINE

"individual scientists to ask better questions," she says.

Among other aspects, the NIH mandate asks researchers to compile a plan that describes the data type that will be generated and shared. They must indicate the software tools needed to work with these data, the applied standards and metadata and the repository where the data and metadata will go. Among the criteria for a repository is that it provide broad, equitable and open access to data and metadata free of charge, that it have unique and persistent identifiers, that it record data origin and chain of custody and that it be sustainable for the long term. One of the most frequent questions she hears, says Basehore, are those related to selecting a data repository. On this issue, she encourages investigators to reach out to colleagues and, within their institutions, to data librarians and grants offices.

In genomics, the repository of choice is often ENA and its sister resources GenBank and DDBJ. The ENA data submission interface has checklists and descriptions about data formats and standards used in different communities. These were co-developed with researchers in those communities, says Cochrane. Submitters get general guidance about submitting data, which might be about oceanic plankton, plant specimens or the human gut microbiome.

Queries arrive at the helpdesk mainly via e-mail, says Cochrane, not phone. "With a dataset arriving every few seconds, you know we would never have enough people to do that job." Help for submitters is also on the EGA site, which is for personally identifiable genetic, phenotypic and clinical data. Access is controlled as it is in the US Database of Genotypes and Phenotypes (dbGaP) and the Japan Genotype-Phenotype Archive. All three can take in data from around the world and distribute it, which makes them a little like core resources, says Freeberg. These repositories are not mirrored the way GenBank, ENA and DDBJ are, but search is linked so users can apply for the access they need. Bioinformaticians staff the helpdesk, and the team can assist users with curating their submissions, she says.



With data sharing, "genomics has done generally better than most other fields," says John Quackenbush.

"I think we can be very proud of the change in the culture in the field," says Juan Antonio Vizcaino.

Vistas in proteomics

Having resources such as the Protein Data Bank and running the Human Genome Project set the stage for data sharing practices in the life sciences, says Juan Antonio Vizcaino, the EBI's team leader in proteomics. EBI resources receive around 500 proteomic datasets a month. Data sharing is something most in the community do "without any complaints," he says, which might also be due to journals requiring data deposition with papers. Except for medical journals, it's rare that proteomics data associated with a paper are not made publicly available. Prepping data to share is work and involves a learning curve, which can be a challenge, especially in smaller labs, he says. The EBI proteomics staff offer training and help scientists with submissions.

Proteomics researchers recognize how important data reuse is, he says, such as for machine learning approaches applied to the proteomics workflow. Reuse to improve biological insight has perhaps had less of an impact, but, for example, he sees teams applying proteogenomics methods to annotate genomic features. He was also part of a project his colleague Pedro Beltrao led in which they built a reference human phosphoproteome and then, using machine learning and 112 datasets from over 6,800 mass spectrometry runs, assigned functional scores to the phospho-sites[1].

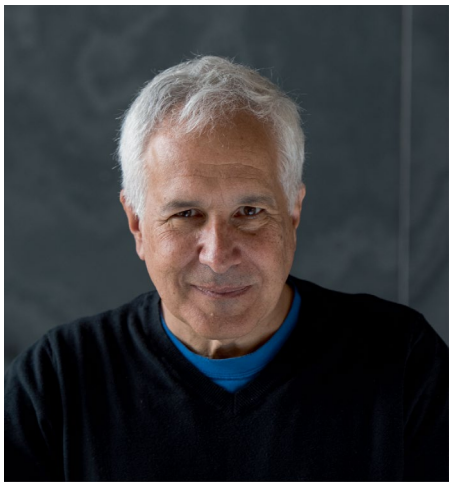
Open data standards enable interoperability between software tools and data repositories and sustainable development of their functionality, says Vizcaino. With mass spectrometry information, it's impossible to support potentially tens of versions of different file formats.

Given the many types of proteomics techniques and instruments, it's difficult to have fully open and interoperable standards. Proteomics used to have multiple XML data formats for raw data output from spectrometers. In 2008 mzML was born as a joint effort between the HUPO Proteomics Standards Initiative, which developed mzData, and Seattle Proteomics Center, which developed mzXML. More stable versions have been developed since then. Along the way, other standards were put forward, but mzML has emerged as an open format that researchers can read and convert their data to.

But mzML is more generic and thus not as optimized as the proprietary raw data formats from vendors, he says. Problems can occur because mass spec files can be big. "But mzML is absolutely necessary, for instance, to compare files between different vendors," he says. The mzML format continues to evolve, but it's hard to support all existing use cases in a single open file format given the many proteomics techniques and diverse instrumentation. The issues with the mzML format applies to other proteomics data types, too, he says, given the many ways researchers are seeking to identify and quantify peptides and proteins.

To enable a more unified standard submission in mass spectrometry as well as to disseminate software analysis pipelines, the proteomics community set up ProteomeXchange. It's for making data deposition and sharing easier across existing resources such as the PRIDE database and the PeptideAtlas. "I think we can be very proud of the change in the culture in the field," says Vizcaino about the last 10 years or so since ProteomeXchange formally started. Seeing the proteomics community adopt data sharing to the degree seen in fields such as transcriptomics "is a big achievement," he says. Proteomics is ahead of some of its subfields. "This is especially clear in the case of metabolomics and sister 'omics approaches such as lipidomics and glycomics."

CREDITS: LEFT: T.H. CHAN HARVARD SCHOOL OF PUBLIC HEALTH; RIGHT: EMBL-EBI



"We share a dataset that's 1,400 terabytes," says Jeff Lichtman.

Of increasing importance, in his view, are restrictions to data sharing for sensitive human proteomics data[2]. Given patient consent issues, laws such as the Health Insurance Portability and Accountability Act (HIPAA) and General Data Protection Regulation (GDPR), and the risk of identifying individuals, controlled access is needed for these data, says Vizcaino. To develop best practices will take infrastructure development for controlled-access sharing and acceptance by the research community without undoing the years of work it has taken to foster an open sharing culture. "This is something that will change some aspects of data sharing in proteomics in the near future," he says. Practices related to issues of law and privacy are common for researchers working with RNA and DNA sequence data, "but they are new for proteomics researchers."

Compared to transcriptomics, says Vizcaino, proteomics needs better quantification standards and better ways to share metadata and experimental protocols. "This is indeed a big gap in the field," he says. A big stumbling block for data reuse is the lack of experimental and biological metadata in public repositories, he says. "This is really where proteomics is behind other fields in terms of data-sharing practices." When ProteomeXchange was started, the objective was to generalize data sharing. With the resources available at the time, metadata collection became secondary, he says.

For effective data reuse, especially in the case of quantitative experiments, says Vizcaino, existing metadata in public repositories will most likely be insufficient. "However, there have been improvements in this area as well, especially thanks to the development of the SDRF-Proteomics format," an increasingly used data standard in the field[3].

Quandaries with imaging data

For researchers using imaging such as confocal microscopy or electron microscopy, data-sharing challenges reach new dimensions as imaging modalities multiply. A number of online resources hold imaging data, and beyond archives there are "added-value databases," which, as Swedlow and his co-authors point out[4], means that datasets are enriched and combined with others through "well-designed analysis, expert curation, and where possible, meta-analysis."

The imaging repositories he finds of note, says Swedlow, include the EMBL/EBI's BioImage Archive, which holds data from all imaging modalities. The data must be associated with a publication or, as the site notes, be "of value beyond a single experiment." The data, he says, have minimal annotation and metadata. The Image Data Resource[5] is another public repository for image datasets associated with a published study and one he and colleagues in the Open Microscopy Environment consortium have built. The datasets are annotated and curated, says Swedlow, and there's a public API for the community to use. Japan's RIKEN hosts the Systems Science of Biological Dynamics database, which provides imaging data to the bioimaging community in Japan.

For smartphone owners, gigabytes are common measurement units, but sharing photos can quickly become a household challenge. Neuroscientist Jeff Lichtman at Harvard University and his colleagues inhabit a petabyte-scale world. "We're already talking about doing datasets that are an exabyte – 1,000 petabytes or a million terabytes," he says. Early in his career, a gigabyte sounded large, he says, but "today's terabyte is last year's gigabyte." Terabytes – a terabyte is 1,000 gigabytes – are common in many labs across neuroscience, genomics, proteomics and imaging.

Lichtman admits his lab sits at an extreme data-producing end, but the amount of information researchers are seeking to share is rising to astronomic heights, he says. "In my case, we share a dataset that's 1,400 terabytes." The dataset comprises annotated electron micrographs that can be interactively explored. Clicks can reveal blood vessels traversing a volume of imaged brain tissue or isolate only those neurons with dendritic spines.

The dataset is based on raw data over a petabyte in size. Few places will host 1,400 terabytes of data for sharing, and it's technically difficult to share that, he says. Along with colleagues at other institutions, the Lichtman lab is part of a project to build a connectomic atlas of the mouse brain. Lichtman embraces the NIH approach that states that, if NIH funds are used to generate data, it must be shared. Sharing is expensive. Like some other labs, Lichtman has an arrangement with Google to host his data to share it with others. But not all labs have these options. "If I had to pay to share 1,000 terabytes of data, it would be way more than an NIH grant just to share it," he says. "I think there is a profound problem right now."

Beyond the ENA's sequence information, EBI holds data such as imaging data and genome assemblies on the scale of tens of petabytes, says Cochrane. Overall, the EBI's data holdings add up to hundreds of petabytes – almost half an exabyte – and that's not counting the raw data streaming off sequencers. Fortunately, says Cochrane, just as data pile up, data reduction techniques are being developed and applied, often before data submission. Network bandwidth is used in various ways to share data more efficiently, and sequence data can be compressed to substantially reduce the data's footprint.

On the subject of data mountains Cochrane and other EBI colleagues stay in touch with Big Data institutions such as the European Organization for Nuclear Research (CERN). There's plenty to talk about, but much is particular to life science data, says Cochrane.

Compared with high energy physics data, the infrastructure for life science data is more distributed. This spread-out infrastructure grows organically and in response to need, he says. In addition to established repositories such as ENA, GenBank and DDBJ, when a biomedical lab or a research community wants to share data, scientists set up a digital resource. Globally, that has led to around 3,000 digital resources that hold biological data, he says. That number is from an ongoing tally just of the resources for which there are associated papers describing them, he says.

The 3,000 figure is from a survey by the Global Bio Data Coalition, an organization with which Cochrane is involved. The alliance is focused on the sustainability of global biodata resources around the world. In a next phase, the team is setting out to characterize these resources. Given these numbers, it's not surprising that, in some subdisciplines of the life sciences, it can be hard to find the right repository home for one's data.

CREDIT: HARVARD UNIVERSITY

Setting standards

The value of data for others extends beyond the paper in which they are first presented, says Lichtman. If and when these datasets are made available, they could generate hundreds of additional papers. "Genomics has sort of figured it out," he says. Once an animal's genome has been sequenced, many scientists can use these data for a wide range of questions. "In my field that is still rare," he says.

To make sense of connectomic data, they have to be partitioned and analyzed, says Lichtman, which is tough work for many. The NIH frequently asks him about the data-sharing standard in connectomics, which, he says, is a great question he can't answer. "There is no norm for connectomics data," he says, "because connectomics is, compared to genomics, in its infancy, so there is no one way to do it." In connectomics, "everything is different about every single data set," he says. Tissues can stem from different animals of differing ages, and sample volumes differ, too. Given the ongoing "explosion of new imaging modalities," he says, such issues are true in connectomics and other fields, too. Transcriptomics has become an imaging modality and delivers a new data type that remains difficult to analyze, and the raw data are challenging to share.

Many studies are now multimodal. Lichtman's electron micrographs can have superimposed functional data generated using molecular labels and with tissue from the same animal. "All of that has to be accessible and interpretable," he says, which means researchers must annotate it to share it. When I mentioned the NIH session at the AACR meeting, Lichtman says it does not surprise him that some researchers wonder "who is going to do the heavy lifting" related to data sharing.

As part of a reaction to size of datasets and the emergence of new data types, Lichtman sees data science emerging as a scientific specialty. Some PhD biologists and postdoctoral fellows add data science to their training. "I think this is a growth area," he says, and an exciting area for junior scientists. Some might not be as interested in the "back-breaking work of doing serial electron microscopy or transcriptomics of many cells simultaneously," he says. "But what do you do with the data once you get it? That is really a bottleneck right now." Without context for the data, data sharing won't work.

Yet to fix

"A data file itself is really pointless," says Freeberg. Data need associated information, which might be about the phenotype or disease or sample quality, and it's crucial to data sharing. In the case of organoids, metadata might include how the organoid was grown or collected. With human data, metadata can include an individual's demographic and health information. "We definitely see this expanding in the human genomic space," she says. When they have new data for a specific dataset, researchers can append the previous data and metadata. Metadata standards emerge from communities. For example, for the Human Cell Atlas, researchers discussed and agreed on what needs to be described in single-cell experiments.

Data need to be shared with associated information, such as metadata, says Mallory Freeberg.

Separately, in collaboration with Global Alliance for Genomics and Health, says Freeberg, she and others are developing better ways to handle and share data that is sensitive to the fact that under-represented communities have, in the past, been harmed by participating in research and clinical trials. "How do we return results in an ethically sound way?" is one of a number of questions researchers in this alliance work on. One project to define how data can be used is Data Use Ontology.

With data classifications, especially about people, it becomes easier to analyze data and draw inferences from them. Classifying people by ethnicity, geographic region or tribe language takes sensitivity. "And there isn't one right way to do this," she says. "There are clearly wrong ways and ways that have harmed people in the past." She and her colleagues want to use lessons learned to foster a standard terminology for the data. It will be built on FAIR principles, which relate to data being findable, accessible, interoperable and reusable, and a second set of principles, the CARE principles, that outline Indigenous data governance. These build on FAIR principles and includes greater control over the way Indigenous data and knowledge are used.

Genetic and genomic data are being generated in a basic research and a health context. For example, newborns are getting their genomes sequenced, and access to these data will vary, says Freeberg. Just because data have been collected doesn't mean they can be deposited in EGA or dbGaP and used in research. This holds true for many projects.

As part of a large-scale undertaking, the UK Biobank is a resource for which genomic, imaging and health data from half a million people in the UK are being collected. Other large-scale projects include the International HundredK+ Cohorts Consortium. As the name indicates, it's about studies involving 100,000 people or more. Big bio data will be getting bigger.

"We have to be honest about how much time it takes," says Freeberg about data management and sharing plans. Software tools help with data prep for sharing and the sharing itself. It takes training for users and potential users, which the EBI offers. "But I think research groups and clinical groups have to be honest and ask for the support," she says. In her assessment, repositories will need more personnel.

"Science as a balancer of the world's inequities is, I think, quite powerful," says Cochrane. The EBI is involved in the CABANA project, which is building capacity and expertise in data science in Central and South America, says Cochrane. The project aims to give greater access to science, data science and eventually infrastructure.

In spite of declarations that data are available or available upon request, the reality can often show a different picture[6,7].

Systems must be built to enable sharing and provide transparency about data use, NCI director Bertagnolli said at the AACR annual meeting press conference. Much data from across the health system still need to be captured, and also data from populations "we are not getting to," she said. She believes it's possible to put in place structures for properly and respectfully sharing data, across all of public health.

In many studies negative data are missing; thus they are also missing in the data-sharing realm. Says Quackenbush, "Even academic negative data, there's no place to publish it." Looking back on her own research in bioinformatics, Freeberg says it frustrated her to

CREDIT: EMBL-EBI

It's understandable that the data-sharing mandate can seem overwhelming for labs not used to data sharing, says Emily Boja.

think the analysis she was doing may have previously been done ten times. But it had not yielded a result and thus was not published. For both wet-lab and computational negative results, "I always personally had this dream of having like a journal of negative results," she says.

"We need to publish negative data," says Elizabeth Jaffee, cancer researcher at Johns Hopkins University and deputy director of the Sidney Kimmel Cancer Center. This necessity applies to work in industry, academia and government. Nobody is rewarded for publishing negative data, but, she says, it would yield lessons and lead to better treatments faster. "Industry has a lot of data associated with new drugs that they don't share but need incentive to do so," she says.

Science is built on falsifiability, says Quackenbush. "That's the acid test: you develop a theory and test it." Data and findings need scrutiny, and from that scrutiny insight emerges. Open science and data sharing play a crucial role in that scrutiny. Back in 2012 he and colleagues in multiple institutions were building a classifer method to predict drug response and looked at the data in two *Nature* papers. These were large-scale studies from the Cancer Cell Line Encyclopedia and the Cancer Genome Project[8,9] that assessed with different methods and platforms how well cancer drugs worked on cell lines, which had been characterized in terms of gene expression and mutations. Of the large group of tested drugs and cell lines, 15 drugs and 471 cell lines were tested in both studies. "When we trained the method on one data set and applied it on the other one, it failed," he says. The gene expression profiles were noisy – it was 2012 – but standards had been applied and the expression profiles correlated well. Correlation was poor between the findings of the two studies in terms of sensitivity to drugs and the associated genomic features of cells.

"There is no way with available data to determine which study is more accurate," Quackenbush and colleagues noted in their paper[10]. "Users of both data sets should be cautious in their interpretation of results derived from their analyses." The day his paper was published, he received calls that he was ruining careers, says Quackenbush.

In the News & Views that accompanied the paper[11], MD Anderson Cancer Center researchers John Weinstein and Philip Lorenzi noted that the findings "sound a note of caution" about interpreting data from such projects but "do not undermine their value." Many variables, they note, can affect the quantitative results from the different pharmacologic assays used. There are "too many differences between cultured cells and patients," especially in regard to the delicate balance between beneficial and toxic effects of cancer drugs. In their view, cell-line pharmacological data are useful to generate hypotheses and to elaborate on other hypotheses, "rather than for formal statistical prediction." They suggested a "joint effort by the teams" to pin down the differences between assays. That would be a way to "support the activities of the many researchers who are using, or will use, these rich data resources." Eventually, the author teams of the two papers worked out reasons for the differing results.

Sometimes, says Freeberg, submitters don't approach EGA until they're ready to publish their paper. "We do what we can," she says, to lobby for a needed "cultural shift." Mandates from the NIH and other funders to set up a data management and sharing plan at the outset of a project are "one way to sort of get the shift in culture to happen."

Swedlow sees room for scientists "to establish themselves as sources of reference datasets." Most importantly, in his view NIH and all other funders "need to support the development of tools and platforms to operate in compliance with the new rules." The Open Microscopy Environment he and his colleagues lead aspires to do some of this for bioimaging, "but there is so much more to do."

Junior scientists are more likely to engage data-sharing concepts and embrace the need to share, "so hearts and minds, plus technology development that engages with junior scientists, would be a great idea," says Swedlow. The NIH move to mandate data sharing by the researchers it funds with submitted data management and sharing plans, "while radical, is correct and timely," he says. "It will be uncomfortable, but it will drive the change that is needed."

Barker would like to see more theoreticians and mathematicians involved in biomedicine, who will bring new ways of looking at data. They can shape the culture of biological research culture itself. A scientist on his or her own usually faces a statistically underpowered dataset, which limits the science that can be done with it. Data collection is necessary, but the journey from terabytes to petabytes and beyond has to involve converting data into information, says Barker. To enable this conversion, more data are needed and context must be brought to the task, she says. "If you can't bring context, it's just entropy; it's just noise."

Indeed, some are unwilling to share data, says Barker. "That's not going to go away until we change the reward system in universities," she says. "You tend to get the behavior you reward." By rewarding scientists for sharing data, it becomes apparent how much value sharing has. "Just turn that little knob, you know, and say, instead of hanging on to your data, let's reward you for sharing it because that's going to change the world."

Vivien Marx ✉
Nature Methods.
✉e-mail: v.marx@us.nature.com

Published online: 28 June 2023

References

1. Ochoa, D. et al. *Nat. Biotechnol.* **38**, 365–373 (2020).
2. Bandeira, N., Deutsch, E. W., Kohlbacher, O., Martens, L. & Vizcaíno, J. A. *Mol. Cell. Proteomics* **20**, 100071 (2021).
3. Dai, C. et al. *Nat. Commun.* **12**, 5854 (2021).
4. Ellenberg, J. et al. *Nat. Methods* **15**, 849–854 (2018).
5. Williams, E. et al. *Nat. Methods* **14**, 775–781 (2017).
6. Gabelica, M., Bojčić, R. & Puljak, L. *J. Clin. Epidemiol.* **150**, 33–41 (2022).
7. Hussey, I. 2023. Preprint at *PsyArXiv* https://doi.org/10.31234/osf.io/jbu9r (2023).
8. Barretina, J. et al. *Nature* **483**, 603–607 (2012).
9. Garnett, M. J. et al. *Nature* **483**, 570–575 (2012).
10. Haibe-Kains, B. et al. *Nature* **504**, 389–393 (2013).
11. Weinstein, J. N. & Lorenzi, P. L. *Nature* **504**, 381–383 (2013).