Mohammed AlQuraishi is a Fellow in Therapeutic Science at Harvard Medical School, Boston, MA 02115, USA.

Peter K. Sorger is Professor of Systems Biology, Head of the Harvard Program in Therapeutic Sciences, and Director of the Laboratory of Systems Biology at Harvard Medical School, Boston, MA 02115, USA. Email: peter_sorger@hms.harvard.edu

Citation:
M. AlQuraishi, P. K. Sorger, Reproducibility will only come with data liberation. *Sci. Transl. Med.* **8**, 339ed7 (2016).

10.1126/scitranslmed.aaf0968

DATA SHARING

Reproducibility will only come with data liberation

IMPROVEMENTS IN HUMAN HEALTH—MADE POSSIBLE BY, FOR EXAMPLE, INNOVATIVE new medicines—are highly dependent on an ecosystem in which academic laboratories publish provocative proof-of-concept studies and in which industrial scientists use these studies to develop therapeutic drugs and test them in the clinic (*1*). So when prominent examples arise of papers in high-visibility journals that cannot be confirmed via independent experiments (often performed by industrial scientists) (*2*), concerns about the reproducibility of biomedical research inevitably arise. In addition to wasting time, effort, and money, irreproducibility threatens the survival of the biomedical research enterprise and the well-being of patients who need better medicines.

It is rarely obvious whether irreproducibility, once identified, is a consequence of biological variability, insufficient sample size and weak statistics, bad reagents, experimental error, or a tendency to base broad claims on narrow evidence (or even, in rare cases, fraud). Such problems could be studied and addressed much more effectively were the data in published papers and figures easily accessible for reanalysis. Alas, this is not the case: The vast majority of preclinical, translational, and clinical research data are locked away in rasterized figures and nontext files that make reanalysis nearly impossible.

As a concrete example, consider Kaplan-Meier curves (survival curves) from clinical trials of anticancer drugs (*3*). Such curves summarize complex and expensive experiments involving hundreds of very ill human patients. However, the numerical data underlying such figures are rarely, if ever, linked to the paper or made available in any other readily accessible location. As a last resort, numerical values can be extracted from rasterized figures by means of image processing, but even then, only a subset of the information can be recovered. Making the data depicted in figures available in a findable, machine-readable format would not be particularly difficult or burdensome (for example, by using the European Molecular Biology Organization's SourceData format). It is now essential, in our opinion, that we transition to a system in which biomedical research data are liberated from dead-end formats and deposited in public repositories as a precondition for public funding and scientific publication.

Liberated data are data in numerical form, findable with text-based or semantic searches, and associated with metadata that contain a description of the data and the way they were collected. Deposition of such data should include, in machine-readable format, information about reagents, experimental protocols, and data processing routines—everything needed to make data interpretable and repeatable (essentially, an enhanced version of existing materials and methods). Liberating data in figures is distinct from, and should precede, the more ambitious aim of the U.S. National Institutes of Health's (NIH's) data science initiative (the Commons) and similar efforts in Europe to make all primary research data findable, accessible, interoperable, and reproducible—the FAIR concept (*4*).

The fact that publication of most crystallographic and genomic data require deposition in public repositories (for example, the Protein Data Bank and the Cancer Genome Atlas) belies the fact that this is not the norm for other types of data; for example, high-resolution multichannel images (collected with million-dollar microscopes purchased with public funds) are usually available only as postage stamp–sized .jpg files. The relatively poor penetration of available tools for publishing primary image data—such as the *Journal of Cell Biology* DataViewer—demonstrates that the absence of suitable standards is not the sole barrier to data dissemination. In fact, there exists surprisingly little consensus that we should take even the first step toward making data accessible. For example, the editors of the *New England Journal of Medicine* recently argued (*5*) that providing source data for figures only encourages "data parasites"—a position that sparked a storm of online protest and a subsequent editorial "clarification." This outburst suggests that most scientists—and presumably most taxpayers—find the *NEJM* position untenable and the information loss it engenders hard to justify.

CREDIT: SAM OGDEN

Nature Genetics chief editor Myles Axton has pointed out that public release of data increases citation rates, and Pardis Sabeti and colleagues have argued that timely release of breakthrough biomedical data is an ethical necessity in the face of serious disease (*6*). Routine public release of biomedical research data will also serve as a brake on rare but egregious cases of fraud and scientific misrepresentation, in much the same way that public disclosure helps to manage conflicts of interest and discourage misconduct. However, the great benefit of findable, machine-readable data is that such data can be reexamined in light of new hypotheses; in fast-moving fields, it seems inevitable that correct data will often be inadequately or incorrectly interpreted at first. Liberated data also can be integrated with other data to make discoveries that are not possible by using a single data set in isolation (the assembly of gene sequences into a chromosome is a self-evident example). Integration of quantitative data promises to address one of the great weaknesses in contemporary molecular biology: a predilection for qualitative "word models" and simple explanatory narratives in the face of clear evidence that normal and disease physiology is controlled by quantitative differences involving many genes.

For example, multiple papers describing screens for host genes involved in HIV and influenza virus infections yielded several gene sets that only partially overlap. Careful meta-analysis showed that these discordant sets were actually consistent with a common molecular model of infection (*7*) when sampling statistics and differences in approach were taken into account. Similarly, a recent genome-wide model of SH2 domain–phosphotyrosine interactions, built with ensemble modeling, integrated multiple nonoverlapping sets of structural and peptide-protein binding data as a means to correct for systematic and random (sampling) errors in the experiments (*8*). The resulting model was more predictive than any single data set or simple consensus representation of the data. As the life sciences become increasingly quantitative and data-rich, the ratio between data generation and data analysis will continue to shift in favor of the latter. By analogy, a recent data set from the Large Hadron Collider's ATLAS and CMS particle physics experiments spawned more than 150 papers in the arXiv repository within a month. The value of publicly available biomedical data will be further increased by programs, currently under way, to make experimental designs and conclusions machine-readable, by using formats such as MIBBI (Minimum Information for Biological and Biomedical Investigations), BioPAX, and Biological Expression Language (BEL).

Why, then, do many biomedical research scientists resist the call to make primary data available in a discoverable and reusable form? In our opinion, it is not that laboratory scientists are selfish or disinterested but, rather, that costs are high and incentives are poor. As part of the NIH-funded LINCS (Library of Integrated Network-based Cellular Signatures) program, our group is mandated to release metadata-tagged primary data from images, multiplex immunoassays, and drug dose-response studies. We have found that data-wrangling tasks such as curation, standardization, normalization, and metadata association account for ~20% of our total effort on the project. More intuitive software tools and better training will reduce this effort, but funding agencies aiming to enhance research reproducibility must acknowledge that it will be far from free. Unfortunately, whereas plans are afoot to pay for data storage and computation costs, there is unlikely to be additional funding for data-wrangling and deposition.

Data liberation is a classic public good in which individuals bear the cost but the community as a whole benefits. Supporting such public good usually requires a carrot-and-stick approach. The stick, in this case, is a mandate on data release at the time of publication (for example, Gene Expression Omnibus deposition for RNA expression profiles).What about the carrot? One way to incentivize deposition and also free up substantial funding is to end the "wasteful tyranny of reviewer experiments" (*9*), which could shift, even slightly, the focus of peer review to the quality and accessibility of existing data, rather than a chase for last-minute experiments that tend to be among the worst designed and least well-validated. Achieving this will require a means to supplement the pool of available reviewers with individuals who have the necessary experience in programming, statistics, and data modeling to review primary data sets; these individuals might logically be students and postdoctoral fellows paid for their work. Prioritizing data quality and accessibility will require changes in the attitudes of reviewers and editors, but innovations in peer review are already under way in a variety of venues. We believe that liberating data at the time of publication as a means to improving reproducibility—arguably one of the most pressing problems in contemporary biomedical science—should not be beyond our grasp as a community.

–Mohammed AlQuraishi and Peter K. Sorger

SUPPLEMENTARY MATERIALS

www.sciencetranslationalmedicine.org/cgi/content/full/8/339/339ed7/DC1
Table S1. Online sources of digital resources.

REFERENCES

1. A. S. Kesselheim, Y. T. Tan, J. Avorn, The roles of academia, rare diseases, and repurposing in the development of the most transformative drugs. *Health Aff. (Millwood)* **34**, 286–293 (2015).
2. M. Baker, Biotech giant publishes failures to confirm high-profile science. *Nature* **530**, 141 (2016).
3. C. Robert, B. Karaszewska, J. Schachter, P. Rutkowski, A. Mackiewicz, D. Stroiakovski, M. Lichinitser, R. Dummer, F. Grange, L. Mortier, V. Chiarion-Sileni, K. Drucis, I. Krajsova, A. Hauschild, P. Lorigan, P. Wolter, G. V. Long, K. Flaherty, P. Nathan, A. Ribas, A. M. Martin, P. Sun, W. Crist, J. Legos, S. D. Rubin, S. M. Little, D. Schadendorf, Improved overall survival in melanoma with combined dabrafenib and trametinib. *N. Engl. J. Med.* **372**, 30–39 (2015).
4. M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J. W. Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S. A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, B. Mons, The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **3**, 160018 (2016).
5. D. L. Longo, J. M. Drazen, Data sharing. *N. Engl. J. Med.* **374**, 276–277 (2016).
6. N. L. Yozwiak, S. F. Schaffner, P. C. Sabeti, Data sharing: Make outbreak research open access. *Nature* **518**, 477–479 (2015).
7. F. D. Bushman, N. Malani, J. Fernandes, I. D'Orso, G. Cagney, T. L. Diamond, H. Zhou, D. J. Hazuda, A. S. Espeseth, R. König, S. Bandyopadhyay, T. Ideker, S. P. Goff, N. J. Krogan, A. D. Frankel, J. A. Young, S. K. Chanda, Host cell factors in HIV replication: Meta-analysis of genome-wide studies. *PLOS Pathog.* **5**, e1000437 (2009).
8. M. AlQuraishi, G. Koytiger, A. Jenney, G. MacBeath, P. K. Sorger, A multiscale statistical mechanical framework integrates biophysical and genomic data to assemble cancer networks. *Nat. Genet.* **46**, 1363–1371 (2014).
9. H. Ploegh, End the wasteful tyranny of reviewer experiments. *Nature* **472**, 391 (2011).

Reproducibility will only come with data liberation
Mohammed AlQuraishi and Peter K. Sorger (May 18, 2016)
Science Translational Medicine **8** (339), 339ed7. [doi: 10.1126/scitranslmed.aaf0968]

Editor's Summary

Science Translational Medicine (print ISSN 1946-6234; online ISSN 1946-6242) is published weekly, except the last week in December, by the American Association for the Advancement of Science, 1200 New York Avenue, NW, Washington, DC 20005.