# Quantifying Biomedical Data Reuse in an Open Science Ecosystem

**Lisa Federer, PhD, MLIS, Data Science and Open Science Librarian**
**Office of Strategic Initiatives, National Library of Medicine**
**National Institutes of Health**

The last decade has seen a significant shift in the ways that academic and research communities think about research data. Data can now be generated more quickly and cheaply than ever before, a phenomenon that is clearly evident in the case of genomic data. The process of sequencing the first human genome, under the auspices of the Human Genome Project, took about thirteen years by the time it was complete in 2003 and cost about $2.7 billion, requiring the collaboration of research institutions from around the world (1). Today, a human genome can be sequenced in about 24 hours at a cost of around $1,000. As a result of such advances not only in the field of genomics, but across the range of research disciplines, the amount of data available today has exploded.

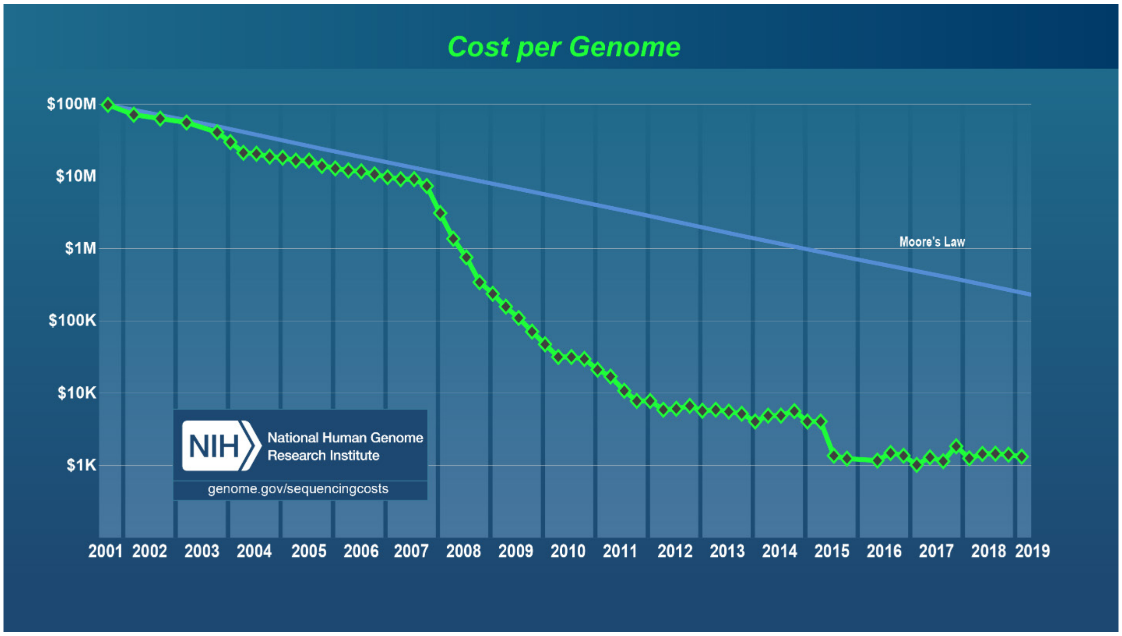

Image source: https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost

Not only do we have more data than ever before, but those data are also increasingly freely available through repositories and other sharing mechanisms. This move toward sharing data has been driven in part by the adoption of policies that require researchers to share their data. PLOS was among the first publishers to adopt such a policy, stating that open access to the literature is only part of making research open, since "without similar access to the data underlying the findings, the article can be of limited use" (2). The International Committee of Medical Journal Editors has instituted a clinical data sharing policy for its member journals, noting that researchers have "an ethical obligation to responsibly share data generated by interventional clinical trials because trial participants have put themselves at risk" (3). Many funders have also adopted such policies, making data sharing a condition for investigators accepting grant funding. At the National Institutes of Health (NIH), a number of policies govern sharing data of different types; in different domains of research, such as clinical data about mental health (4); specific research initiatives, such as the Human Connectome Project (5); or funding mechanisms (6), with plans underway for an overarching policy on data management and sharing that will apply to all NIH-funded research (7,8).

Not all investigators share their data simply because they are required to do so; a growing number of researchers have adopted sharing practices as part of a cultural shift towards open science, a trend in which research products are made openly available. The move toward open access publications is one part of this trend, but open science encompasses digital objects from across the entire research life cycle, including data and code. Enabling access to research products is seen as a way to "foster equality, widen participation, and increase productivity and innovation in science" (9). In light of recent concerns about irreproducible research, open science practices are beneficial to increasing transparency and thereby enhancing research reproducibility (10).

As a result of these advances in technology and changes in science policy and culture, researchers have a wealth of public data available to them. The National Library of Medicine (NLM) plays a significant role in this data sharing ecosystem. As the world's largest biomedical library, NLM not only houses a comprehensive collection of literature, but also provides access to a wide range of biomedical data through a number of databases administered by NLM's National Center of Biotechnology Information (NCBI). Each day, NLM sends out over 115 terabytes of data to 5 million users, as well as adding to its own data holdings by receiving over 15 terabytes of data from around 3,000 users. While a significant source of biomedical research data, NLM is only one part of the big picture for data sharing. NIH alone hosts or funds over 80 domain-specific repositories, housing data related to a specific disease, of a specific type, or funded by a specific institute of the NIH (11). As of this writing, the Registry of Research Data Repositories (re-3data) lists over 1,200 repositories collecting data related to the life sciences (12). Add to this the many institutional repositories that house their investigators' data, as well as generalist repositories, such as Mendeley Data, Zenodo, Dryad, and figshare, that accept a range of data types from various disciplines, and it becomes clear that the universe of publicly available biomedical research data is vast.

Despite all the time, effort, and funding that has been put into making research data publicly available, a fundamental question nonetheless remains relatively unanswered: what happens to all of these datasets? In theory, reusing existing data rather than collecting more yields many benefits to science and society on the whole. Reusing data increases the return on investment of the original funding by yielding additional discoveries and knowledge, as well as saving funds that would have been spent on collecting new, potentially duplicative, data. The time for translation from research findings to life-saving clinical applications may be sped up by reusing existing data rather than taking additional time to collect new. Making data publicly available can also help democratize the practice of science, enabling researchers who may not have access to large amounts of funding or expensive laboratory technology to nonetheless contribute to knowledge creation.

Understanding data reuse can also pave the way for meaningfully rewarding researchers who share their data. Being able to reward researchers who share might also make data sharing more appealing for researchers who are not necessarily open data enthusiasts. Some researchers consider data sharing a burden or worry that making their data available opens them up to being "scooped," concerns that might be mitigated by providing credit for researchers who share data. Science is, after all, a credit economy; if a research team wants to build on my ideas, they need not pay me to do so, but instead give me credit by citing my article in their publication. While a citation in and of itself has no actual monetary value, it indirectly has very real value as a means for demonstrating a researcher's scientific productivity and impact, which in turn form the basis for career advancement in the form of professional recognition, tenure, promotion, and funding. Citations to articles, though an imperfect measure, are a method to quantify the difficult-to-define concept of scientific impact. However, such measures privilege journal articles as the only research output meriting reward, when in fact other research outputs, such as data or code, can have meaningful impact.

A move toward rewarding data sharing is in large part a culture change that must be driven by stakeholders involved in scientific reward, particularly funders and institutions. Indeed, several major funders, including the National Science Foundation (NSF) and NIH have already formally recognized datasets as research products that can be reported to demonstrate researchers' scientific impact for grant applications as well as progress toward grant aims on progress reports (13,14). Some institutions likewise have begun to consider data and other research products in considerations of researchers' scientific output and impact; the Montreal Neurological Institute (MNI), for example, has adopted an institution-wide open science policy that recognizes shared data as a research output in the tenure and promotion process (15).

However, technological challenges remain that hinder efforts to reward data sharing. Using article citations as a means of quantifying impact works because we have well-established mechanisms for tracking such citations. While the exact citation style may differ from one journal to another, authors generally understand how and where to cite an article, and journals know how to appropriately tag citations to enable them to be tracked by systems that capture citations. The same is not true for data; while groups like FORCE11, CODATA, and the International Council for Scientific and Technical Information (ICSTI) have made efforts to help standardize data citation (16,17), uptake among authors and publishers remains relatively low. In fact, some debate remains about whether data citations are even the most appropriate way to acknowledge the contribution of shared data. Some authors choose to recognize data creators in the article's acknowledgement, and some data creators have argued that they should be co-authors on any papers that arise from secondary analysis of their shared data, although sharing data alone does not satisfy the authorship criteria outlined by the ICMJE (18). In the absence of a widely-adopted standard for citing data reuse, quantifying data reuse is impossible in practice, so even the adoption of policies that reward data sharing will be difficult to implement.

In the absence of a reliable means to quantify data reuse, it is still worthwhile to consider how we will eventually reward data sharing at some point in the future. Careful consideration of the meaning of data's value and impact may help avoid some of the perverse incentives that have arisen as a result of the ways that bibliometrics are used to measure the impact of articles by citations (19). One issue is that not all citations to an article mean the same thing, yet all are counted equally when it comes to measuring impact by citation count. Eugene Garfield enumerated fifteen reasons for citing articles, not all of them positive, including "criticizing previous work" and "disclaiming work or ideas of others" (20). For example, the paper in which Andrew Wakefield incorrectly connected autism to vaccinations has been relatively highly cited, with 184 citations according to Google Scholar, 76 citations according to Web of Science, and 74 citations according to Scopus. The disparity in citation counts across various platforms presents its own complication, but also problematic is that most of these citations are in the context of articles that discredit his findings, and simple citation counts would not be able to distinguish this article from another that has been cited a similar number of times.

Similarly, not all instances of data reuse are identical. In genomics research, it is a common practice to pool multiple datasets from different studies and different researchers to achieve adequate statistical power, and the standardization of this type of data means it is possible to do so, since data from multiple sources will be largely interoperable (21). Clinical data, on the other hand, is far less standardized, with researchers often recording the same concept using different terminology or phrasing questions to patients in slightly different ways that mean it is often infeasible to combine clinical datasets even if they are on similar topics (22). If a researcher creates a dataset that is reused as one of several hundred combined together in a genomic study, should that reuse be counted the same as a clinical dataset that is used on its own to entirely form the basis for a new study? Datasets themselves may also have different value based on their contents as well as varied potential for reuse. For example, compare a dataset collected from patients with an extremely rare disease and a dataset collected from patients with heart disease, the most common cause of death in the United States. A dataset on a common condition with high disease burden will almost certainly be reused more than one that covers a rare, and therefore likely less-researched disease. However, it could be argued that the rare disease dataset has greater value since it would be more difficult to re-collect such data than it would be to re-collect data on heart disease. Relying simply on counts of data citation makes it difficult to meaningfully reward researchers in ways that recognize the complexity of data collection and research.

As we move toward a future that is not far off when data reuse can be feasibly tracked and quantified, it will be important for institutions, funders, and other stakeholders to think about how to incorporate metrics for reuse into the scientific system of credit and reward. Overlooking data as an important research output that merits its own recognition and reward means we risk disincentivizing sharing. On the other hand, oversimplifying the practice of rewarding data creators for reuse means we risk creating some of the perverse incentives that have arisen from bibliometrics and led to undesirable research practices like excessive self-citation. It is therefore worth careful consideration now of how we can create a reward system that meaningfully recognizes the place of shared data in the research ecosystem.

**References**

1. National Human Genome Research Institute. The Human Genome Project FAQ [Internet]. 2018 [cited 2019 Sep 4]. Available from: https://www.genome.gov/human-genome-project/Completion-FAQ
2. Silva L. PLOS' new data policy: Public access to data [Internet]. EveryONE: PLOS ONE Community Blog. 2014 [cited 2017 Apr 2]. Available from: http://blogs.plos.org/everyone/2014/02/24/plos-new-data-policy-public-access-data-2/
3. Taichman DB, Sahni P, Pinborg A, Peiperl L, Laine C, James A, et al. Data sharing statements for clinical trials: A requirement of the International Committee of Medical Journal Editors. N Engl J Med [Internet]. 2017;376(23):2277–9. Available from: http://www.nejm.org/doi/10.1056/NEJMe1705439
4. National Institute of Mental Health. NOT-MH-15-012: Data Sharing Expectations for Clinical Research Funded by NIMH [Internet]. 2015 [cited 2019 Sep 5]. Available from: https://grants.nih.gov/grants/guide/notice-files/not-mh-15-012.html
5. National Institutes of Health. RFA-MH-10-020: The Human Connectome Project (U54) [Internet]. 2009 [cited 2019 Sep 5]. Available from: https://grants.nih.gov/grants/guide/rfa-files/rfa-mh-10-020.html
6. Trans-NIH BioMedical Informatics Coordinating Committee (BMIC). NIH Data Sharing Policies [Internet]. U.S. National Library of Medicine; 2019 [cited 2019 Sep 4]. Available from: https://www.nlm.nih.gov/NIHbmic/nih_data_sharing_policies.html
7. National Institutes of Health. NOT-OD-19-014: Request for Information (RFI) on proposed provisions for a draft data management and sharing policy for NIH funded or supported research [Internet]. 2018 [cited 2018 Nov 11]. Available from: https://grants.nih.gov/grants/guide/notice-files/NOT-OD-19-014.html
8. National Institutes of Health. National Institutes of Health Plan for Increasing Access to Scientific Publications and Digital Scientific Data from NIH Funded Scientific Research [Internet]. 2015 [cited 2017 Jul 19]. Available from: https://grants.nih.gov/grants/NIH-Public-Access-Plan.pdf
9. Levin N, Leonelli S, Weckowska D, Castle D, Dupré J. How Do Scientists Define Openness? Exploring the Relationship Between Open Science Policies and Research Practice. Bull Sci Technol Soc [Internet]. 2016;36(2):128–41. Available from: http://journals.sagepub.com/doi/10.1177/0270467616668760
10. Shrout PE, Rodgers JL. Psychology, Science, and Knowledge Construction: Broadening Perspectives from the Replication Crisis. Annu Rev Psychol [Internet]. 2017; Available from: http://www.annualreviews.org/doi/abs/10.1146/annurev-psych-122216-011845
11. Trans-NIH BioMedical Informatics Coordinating Committee (BMIC). NIH Data Sharing Repositories [Internet]. U.S. National Library of Medicine; 2019 [cited 2019 Sep 4]. Available from: https://www.nlm.nih.gov/NIHbmic/nih_data_sharing_repositories.html

12. re3data.org. Life Sciences Repositories [Internet]. 2019 [cited 2019 Sep 4]. Available from: https://www.re3data.org/search?subjects[]=2 Life Sciences
13. National Institutes of Health. NIH and Other PHS Agency Research Performance Progress Report ( RPPR ) Instruction Guide [Internet]. 2017. Available from: https://grants.nih.gov/grants/rppr/rppr_instruction_guide.pdf
14. National Science Foundation. Dissemination and sharing of research results [Internet]. Vol. 2012. 2010. Available from: http://www.nsf.gov/bfa/dias/policy/dmp.jsp
15. Ali-Khan SE, Jean A, MacDonald E, Gold ER. Defining success in open science. MNI Open Res [Internet]. 2018 [cited 2018 Nov 11];2:2. Available from: http://www.ncbi.nlm.nih.gov/pubmed/29553146
16. Data Citation Synthesis Group. Joint declaration of data citation principles [Internet]. Martone M, editor. FORCE11; 2014 [cited 2017 May 19]. Available from: https://www.force11.org/group/joint-declaration-data-citation-principles-final
17. CODATA. CODATA-ICSTI Data Citation Standards and Practices [Internet]. [cited 2019 Jan 3]. Available from: http://www.codata.org/task-groups/data-citation--standards-and-practices
18. International Committee of Medical Journal Editors. Defining the Role of Authors and Contributors [Internet]. 2017 [cited 2017 May 19]. Available from: http://www.icmje.org/recommendations/browse/roles-and-responsibilities/defining-the-role-of-authors-and-contributors.html
19. Stephan P. Research efficiency: Perverse incentives. Nature. 2012;484(7392):29–30.
20. Garfield E. Can citation indexing be automated? In: Statistical Assocation Methods for Mechanized Documentation. 1964. p. 84–90.
21. Hong EP, Park JW. Sample size and statistical power calculation in genetic association studies. Genomics Inform [Internet]. 2012 Jun [cited 2018 Apr 1];10(2):117–22. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23105939
22. Richesson RL, Nadkarni P. Data standards for clinical research data collection forms: Current status and challenges. J Am Med Informatics Assoc. 2011;18(3):341–6.