SOUNDING BOARD

Data Authorship as an Incentive to Data Sharing

Barbara E. Bierer, M.D., Mercè Crosas, Ph.D., and Heather H. Pierce, J.D., M.P.H.

Data from well-designed and well-executed research not only are useful for the original purpose and secondary analyses by the original researchers but also can be repurposed for a variety of applications, including independent replication, avoidance of duplicative studies, generation or testing of new hypotheses, and the general advancement of clinical and biologic understanding. No longer a hypothetical or occasional occurrence, the use of research data by persons other than those who originally gathered the data, termed "data sharing," is currently encouraged or mandated by parallel efforts in the legislature through the 21st Century Cures Act, biomedical journal leadership through the draft data-sharing policy of the International Committee of Medical Journal Editors, charitable foundations such as the Wellcome Trust and the Bill and Melinda Gates Foundation, and the National Institutes of Health (NIH) in its recent request for information on data management and sharing strategies. Data sharing, whether elective or required, creates an obligation for the original investigators who obtain funding, design studies, collect and analyze data, and publish results to make their curated data and associated metadata available to third parties. Despite the major effort that data collection may have taken, sometimes work that continues over the course of decades, there is rarely academic recognition or reward for data sharing itself.

We believe that both as a matter of fairness and as a matter of providing an incentive for data sharing, the persons who initially gathered the data should receive appropriate and standardized credit that can be used for academic advancement, for grant applications, and in broader situations. We propose a system of recognition whereby data generators are identified and cited by means of a designation that would be standardized and differentiated from the designation of the authors of a peer-reviewed journal article.

Although it has been recognized that appropriate and meaningful incentives are essential to capitalize on the promise of data sharing[1] and that crediting data generators is key in this effort,[2] to date there has been no systematic implementation of a standard process and method to credit original data generators. Principles of good data management, such as curation, data citation, and archiving, have been proffered[3,4] but have not translated into a comprehensive model that can be adopted across multiple stakeholders, from academic institutions to journals. Indeed, the current system of academic promotion depends heavily on the number of original peer-reviewed publications, the impact factor of the journal in which the work is published, and the relative placement of the author in the citation. This system discourages data sharing by creating incentives for investigators to maximize the publication of subsequent analyses from a given data set without competition. Some have even claimed that the adoption of a standard of data sharing would impede important research.[5]

An important question, then, is how best to modify systems of apportioning academic credit to better align incentives for data sharing with the advancement of science and medicine. Collaboration with the primary data generators is encouraged and can overcome barriers to data sharing, but in practice this approach may be cumbersome or infeasible and might discourage interpretations that question the original design or that disagree with the hypotheses of the originating investigators. Conversely, those who gathered the original data and wrote the primary manuscripts should not be presumed to have been involved with, vouch for, or agree with every subsequent publication that analyzes the original data set. Data-sharing mandates as a condition of grant award or publication do not address issues of fairness to, or incentives for, the primary data generators and could result in compliance-focused sharing, in which investigators expend the minimal effort and resources to meet the requirements but do not create a data set with the necessary metadata or customized analytic tools to allow for robust future publications.

A meaningful and standardized designation for persons who contribute the data that are used in a peer-reviewed publication must reflect two points: first, the significance of the contribution to the scholarship, whether as part of a primary or secondary analysis; and second, the independence of the data generator from analyses performed by data users that lead to a subsequent publication when the resulting article does not derive from a collaboration. Each of these two potentially conflicting roles must be represented in the choice of what to title this designation. Terms such as "data curator," "data custodian," and "data steward" may clarify the distinction from an author but may not sufficiently capture the importance of the role. For the purposes of this proposal, we use the term "data author" as a placeholder for the designation, but we realize that this term could be confused with "author" in a publication or curriculum vitae. Furthermore, we prefer the term "data author" to "data generator" or "data contributor" since there may be many persons who contribute to the generation of data, but not all those persons will assume responsibility for the acquisition, curation, deposition, and integrity of the data and associated metadata.

We propose that, in order to be cited as a data author, a person must have made substantial contributions to the original acquisition, quality control, and curation of the data, be accountable for all aspects of the accuracy and integrity of the data provided, and ensure that the available data set follows FAIR Guiding Principles, which instruct that the data and metadata meet criteria of findability, accessibility, interoperability, and reusability.[3] Data authors are responsible for the integrity of the data set but are not responsible for the scientific or clinical conclusions of the analyses drawn from the data unless they were also listed as authors of the original manuscript. This distinction would permit healthy disagreements while acknowledging the use of a data set. We anticipate that a person could be designated as both an author and a data author on a single publication, whether through the generation, curation, and analysis of the data set in the first publication or in a secondary analysis by either the original investigators or in collaboration with other authors. A given manuscript could have distinct data authors and authors whose primary contribution has been to perform data analysis of an existing data set. Since authors who use the data of others can vouch only for the analysis of those data, and not for the collection or veracity of the data, an open question is whether the criteria for authorship should be refined to include or to differentiate between authors whose sole contribution has been at the level of data analysis and authors who are also data authors. A number of possible scenarios are shown in Figure 1.

In the context of guidelines established by granting agencies and journals, the persons involved in acquiring and collating the data would probably be responsible for determining who meets the criteria for data authorship. Typically, these decisions will be made when the data set is developed. The listing of data authors would be transmitted with the data set at the time it is placed in a repository for access by third parties. The establishment and wide adoption of data-authorship criteria might address the fact that studies that fail to reject the null hypothesis, especially clinical trials, are less often accepted for publication. Such measures may also result in the promotion of data journals (i.e., peer-reviewed, open-access journals that describe data sets, software, models, and databases)[6] and may capacitate deposition into data repositories that directly generate a data citation (e.g., Harvard Dataverse, Dryad, and domain-specific repositories).

For such a system to gain traction, data authors would need to be listed in the primary publication, on publication in data journals, or with direct citation from data repositories; cited in Medline as data authors; and be searchable in the National Library of Medicine and similar database resources (e.g., bioCADDIE). Over time, high-quality, usable data sets are likely to be cited more commonly, and the number of citations would be reflected on a person's curriculum vitae. Academic institutions could modify their faculty reporting formats and promotion criteria to recognize the contributions that are indicated by data authorship. The instructions for the academic narrative, such as in the current NIH biosketch,[7] should be modified to allow the inclusion of publications that cite data authorship, indicating the number of citations for each

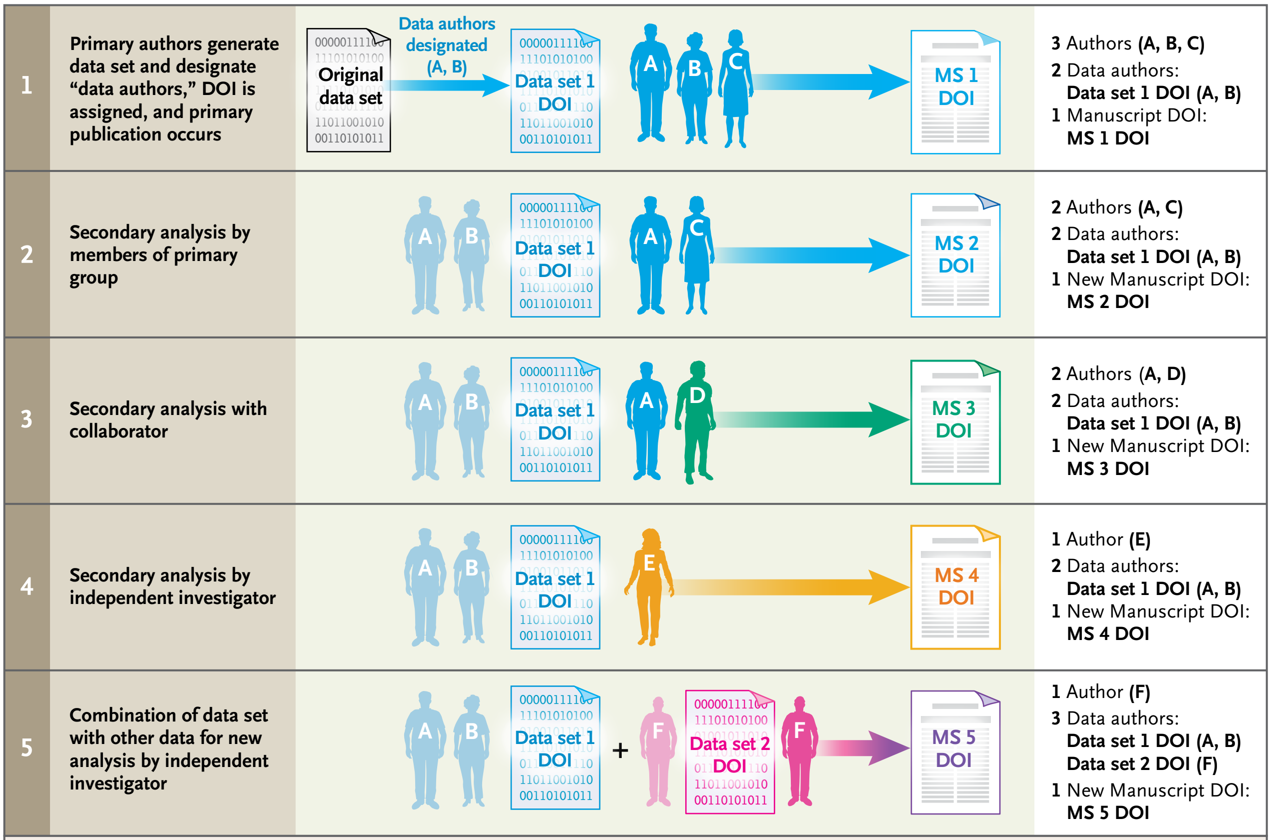

Figure 1. Credit for Data Sharing and Tracing the Data Set.

An individual researcher (indicated by letters A through F) may be designated and credited as an author, a data author, or both, depending on the person's contribution to the data and analysis in the published work. DOI denotes digital object identifier, and MS manuscript.

contribution and an explanation of its significance. Similarly, nothing would prevent investigators from including data authorship as one of their substantive contributions for consideration in promotions and in grant applications. The development of and methods for a "d-index" metric for data authors, similar to the "h-index" or "i10-Index" for authors, as an attempt to measure both productivity and the importance of the contributions of data authors, might further propel data-sharing efforts. Journal policies, peer review, and editorial practices would need to evolve to include data authorship. Granting agencies and foundations could consider data authorship, as well as contributions to data sharing generally, to be an element of review for further funding; could have a means to track data-sharing requirements for funding, if any; and could monitor the subsequent use of data as an additional surrogate for importance, significance, and effect of the original, funded research.

We acknowledge that there are many unanswered questions. The affirmative standards and responsibilities for the integrity and curation of the data set may need to be further elucidated. Should the designation be "data author," or is there a more appropriate term? Should there be a hierarchy of designations to define multiple types of contributions (e.g., data curator, data analyst, and data statistician), as has been suggested previously?[2] Should primary data generators be informed or consulted about upcoming manuscript submissions on which they will be cited as data authors, and if so, how? Should a data generator be able to decline being cited as a data author if the conclusions of the publication could harm the reputation of the primary investigators? How does the designation of a limited number of persons as data authors align with existing data-citation principles that state that data citation should facilitate giving credit and attribution to all contributors to the data? What modification or transformation of a data set warrants a new digital object identifier (DOI) or other machine-readable persistent unique identifier? And does a transformed data set, with its new DOI, carry metadata from the original citation, and is that original metadata obligatory (i.e., should provenance metadata be included as part of the required citation metadata of a data set)? Should standards exist for a comprehensive data set, or is the creation of multiple small data sets, each with its own DOI, appropriate? Should there be a system for monitoring or auditing designation of data authors?

This proposal highlights the need to link the extensive efforts to date to define, standardize, and implement citation formats and principles with systems for credit and attribution that have not been modified as data sharing has become more prevalent. We appreciate that the promulgation and acceptance of citations for data generation will take time, including time for the National Library of Medicine to index the designation, time for investigators to use data sets in secondary analytics and to cite those data sets using the DOIs, and time for data citations to develop. But these metrics are possible only after the principles are framed, endorsed, broadly adopted, and consistently applied. We think that it is time for the international research and academic communities, industry, funders, and journals to take the next steps to answer these questions and ensure that credit for data sharing is keeping pace with calls to increase access to data. Further development of the concept of and criteria for recognition of the contributions of data generators is timely and will propel data sharing for the advancement of science and public health.

Disclosure forms provided by the authors are available with the full text of this article at NEJM.org.

From the Multi-Regional Clinical Trials Center of Brigham and Women's Hospital and Harvard University and the Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston (B.E.B.), and the Institute for Quantitative Social Science, Harvard University, Cambridge (M.C.) — all in Massachusetts; and the Association of American Medical Colleges, Washington DC (H.H.P.).

This article was published on March 29, 2017, at NEJM.org.

1. Lo B, DeMets DL. Incentives for clinical trialists to share data. N Engl J Med 2016;375:1112-5.
2. Kalager M, Adami H-O, Bretthauer M. Recognizing data generation. N Engl J Med 2016;374:1898.
3. Data Citation Synthesis Group. Joint declaration of data citation principles — final. San Diego, CA: FORCE11, 2014 (https://www.force11.org/group/joint-declaration-data-citation-principles-final).
4. Wilkinson MD, Dumontier M, Aalbersberg IJ, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 2016;3:160018.
5. The International Consortium of Investigators for Fairness in Trial Data Sharing. Toward fairness in data sharing. N Engl J Med 2016;375:405-7.
6. Akers K. A growing list of data journals. Ann Arbor: University of Michigan Library, May 9, 2014 (https://mlibrarydata.wordpress.com/2014/05/09/data-journals/).
7. New biographical sketch format required for NIH and AHRQ grant applications submitted for due dates on or after May 25, 2015. National Institutes of Health and Agency for Healthcare Research and Quality, December 5, 2014 (https://grants.nih.gov/grants/guide/notice-files/NOT-OD-15-032.html).

DOI: 10.1056/NEJMsb1616595