Trials and Tribulations—11 Reasons Why We Need to Promote Clinical Trials Data Sharing

Atul J. Butte, MD, PhD

With the release of the sharing clinical trial data consensus study report by the National Academy of Medicine in 2015,1 the International Committee of Medical Journal Editors (ICMJE) launched an effort that eventually led to requiring authors to submit data sharing statements with their manuscripts starting in mid-2018. Data sharing statements are meant to show readers how to access the deidentified individual-participant data (IPD) that underlie the results shown in a manuscript. Danchev et al2 explored how successful that effort has been.

Starting with 486 clinical trial articles published in JAMA, New England Journal of Medicine, and The Lancet that contained a data sharing statement (or were required to have one), Danchev et al2 found that only 2 deidentified IPD data sets were publicly available on a journal website (both were associated with the same trial), and only 17 were provided in a secure repository, totaling just 4% of the articles. Funding for articles without data sharing statements came from both the National Institutes of Health (NIH) and non-NIH sources. Although many authors stated that data would be made available in public or private repositories or conditionally by request, most of these data sets never end up being available to the public.

Raw research data have been successfully shared by thousands of scientists for decades. The launch of GenBank in the 1970s (along with other similar repositories worldwide) led to a new standard level of openness with researchers sharing their raw data with others. Many journals now require authors and many funding agencies expect their grantees to deposit raw research data into these repositories. Regulatory agencies, such as the US Patent and Trademark Office, started to required sequence data deposition in the 1990s. Similar repositories for the sharing and dissemination of other molecular or cellular data now exist; the yearly database issue of Nucleic Acids Research serves as an informal catalog of these and currently lists more than 1600 different databases.

Then why is it so hard to get clinical researchers to share their research data? Danchev et al2 list several challenges to sharing IPD, including risks to patients (privacy and confidentiality), inappropriate use of data, and loss of potential secondary research output. Although some academic researchers may fear that others will potentially make the key discovery in their data set that they missed, industry researchers might fear rogue analysts who add confusion to regulatory review. Although the NIH has recently updated and formalized their data sharing policy for grantees, those policies could still be made stricter,3 in that just requiring a data sharing statement in a publication or a data management and sharing plan in a grant proposal does not ensure that data are actually openly shared.

If the sticks are only slowly getting sharper, are the carrots getting sweeter? As the principal investigator of ImmPort, a repository of clinical trials data, I have seen many trialists successfully deposit their raw deidentified IPD for unrestricted dissemination and reuse.4 Here, I briefly provide 11 positive reasons that make the case for sharing IPD from clinical trials.

The first 5 reasons address researcher users themselves. The lack of available research data is a well-known reason that impacts the reproducibility of complex scientific experiments.5 If research is currently facing a reproducibility crisis, perhaps the easiest way to correct some of this problem is by releasing raw data. Separately, lessons from fraudulent research over the past decade has shown that as more diagnostics and therapeutics are derived from complex combinations of data elements,
it becomes more important that the scientific reasoning behind the selection of these variables be openly shared to improve transparency, including the raw data. Released raw research data could help subsequent researchers gain visibility into failed trials,^6 potentially improving their next study, which is especially important given the sacrifices made by the research participants.

For successful trials, researchers could potentially use raw research results to find precision subsets that potentially benefit the most or least from an intervention, leading to more directed research.7 As more successful trials lead to approved products for use in medical care, raw clinical trials data could enable researchers to perform digital comparative effectiveness across studies or directly compared against electronic health record–based real-world data.

The next 2 reasons are for research participants. Releasing data could force the hand of trialists to speed results reporting,^8 whereas returning data to the participants could help disease-oriented groups to design and run their own clinical trials.9

Three reasons are for nonscientist data users. Data from sophisticated clinical trials, especially those with centralized expert cores evaluating patients, could be used to enable learning. For example, a core facility looking at kidney biopsy specimens for a large multicenter trial and evaluating them for pathologic features could be creating a data set invaluable for training future pathologists. Raw research data could also be used to enable companies and ventures. NextBio, Pathwork Diagnostics, and NuMedii are 3 of many companies that have created products and services starting with public data. Of importance, clinical trials data could support public policies. For example, as governments worldwide purchase therapeutics and vaccines to address the coronavirus disease 2019 pandemic, it could be argued that all the raw clinical trials data about these products should be made publicly available.

The final reason is for journals to address new attacks on trust and believability. With the increase in use of social media and in potentially viral online mobs attacking articles, it is not yet clear how journals are going to respond to the increase in both amateur and professional skeptics. We have even already seen on Twitter a journal editor criticizing the approval of a clinical trial article by another journal when that article had a majority of authors from a pharmaceutical company. How does a journal respond? How will journals respond to health systems that will now look at drug effectiveness in real-world clinical data? Will the data match? How will journals respond to payers incentivized to challenge expensive drug approvals and who will now want to see the raw data with billions of dollars dependent on these findings? How will journals counter when government officials label them as providing fake news? The only way journals can address these new attacks on science is to require more of authors; it is now finally in their interest to require IPD data release.

If clinical trial data sharing continues to stall despite new requirements by the NIH and ICMJE, I call on these agencies to increase their commitment and enhance their requirements. The case for sharing has already been made; thousands of scientists have shown they can openly share their data. It is time for any remaining limiting exceptions to be exceptionally limited.
Institute, Gerson Lehman Group, AlphaSights, Covance, Novartis, Genentech, Merck, and Roche; being a shareholder in Personalis and NuMedici; being a minor shareholder in Apple, Facebook, Alphabet (Google), Microsoft, Amazon, 10x Genomics, Illumina, CVS, Nuna Health, Assay Depot, Vet24seven, Regeneron, Sanofi, Royalty Pharma, Twist Bioscience, Pacific Biosciences, Editas Medicine, Invitae, AstraZeneca, Moderna, Biogen, Paraxel, and Sutro; receiving honoraria and travel reimbursement from Johnson & Johnson, Roche, Genentech, Pfizer, Merck and Co, Eli Lilly and Company, Takeda, Varian, Mars, Siemens, Optum, Abbott, Celgene, AstraZeneca, AbbVie, Westat, the American Medical Informatics Association, American Society for Clinical Investigation, AcademyHealth, American Association of Allergy Asthma and Immunology, American Medical Association, American Medical School Pediatric Department Chairs, American Urological Association, Asia America MultiTechnology Association, Association for Academic Health Sciences Libraries, Association for American Medical Colleges, Autodesk, CTIC, California Office of Planning and Research, Children’s Hospital Boston, Dana Farber Cancer Institute, Detroit International Research and Educational Foundation, FASEB, FH Foundation, Flare Capital, Georgetown, Helix, Hudson Alpha, Human Proteome Organization, International Society for Advancement of Cytometry, Kneed Media, Mars, Mayfield, Microsoft, National Academies, National Institute of Child Health and Human Development, Optum Labs, Precision Medicine World Conference, Rady Children’s Hospital, Regenstrief Institute, Rock Health, Samsung, Scripps Translational Science Institute, Stanford University, Tensegrity, The Transplantation Society, Three Lakes Partners, Translational Bioinformatics Conference, United Network for Organ Sharing, University of Arkansas, University of Chicago, University of Kentucky, University of Michigan, University of Pennsylvania, University of Virginia, Washington University in Saint Louis, WuXi; receiving royalty payments through Stanford University for several patents and other disclosures licensed to NuMedici and Personalis; and receiving funding from the National Institutes of Health, Northrup Grumman, Genentech, Johnson & Johnson, the US Food and Drug Administration, the Robert Wood Johnson Foundation, the Leon Lowenstein Foundation, the Intervennial Foundation, Priscilla Chan and Mark Zuckerberg, the Barbara and Gerson Bakar Foundation, the California Governor’s Office of Planning and Research, the California Institute for Regenerative Medicine, L’Oreal, and Progenity.

REFERENCES


