Commentary

What Is Data Sharing and Why Should Biomedical Researchers Embrace it?

Martin Bobrow[1]

The idea that scientific data should be shared with other scientists is not new. Basic scientific method demands that published results should be amenable to checking and replication by others, so enough data to enable this should always have been made available after publication (although that has not always been the case in practice). In some physical science disciplines, such as astronomy and some branches of physics, where large observational datasets are gathered at great expense, there have long been arrangements for many scientists to work on the same data sets. Recently, however, the concept has become fashionable in the biomedical sciences. Fashions always deserve to be treated with caution, so what underlies this particular movement at this time?

As pointed out by Hutchinson,[1] making data available to others supports checking and replicating findings, allows novel interpretations, and saves the cost of acquiring similar data again—although of course there are important reasons for doing completely independent replications of some studies. Collecting some sorts of medical research data, such as those involving genomics and larger epidemiological studies, has become expensive. The huge amounts of data collected are usually capable of being exploited for addressing different types of question, and are sometimes larger than the group collecting the data could fully exploit in a reasonable period of time. People (including those who donate their personal information or samples for study) want to go beyond first analyses toward getting clinically useful information out, validated, and in practice as soon as possible and are more interested in this than in the career issues of individual researchers. Research funding agencies want results and do not want to fund the same data gathering several times, so many have adopted policies encouraging data sharing. All of these combine to give momentum to the idea of data sharing. As also pointed out by Hutchinson, the advent of the internet and sophisticated bioinformatics has made it feasible to share primary data in a way that would have been impracticable some while ago.

For many of us, this move started with the Human Genome Project, whose leaders decided at the Bermuda conferences[2] to put all primary DNA sequence directly and immediately into the public domain, available to all including their direct competitors. This move, making the sequence a public property available to anyone who wanted it, was transformational and kick-started the massive development of biological research of the past 20 years. These concepts of data as essential infrastructure, of projects that were less about testing hypotheses than about accumulating essential generalized information, were further elaborated in conferences at Fort Lauderdale[3] and Toronto.[4] By chance, and not for exactly the same reasons, the world of clinical trails has also very recently seen a huge cultural move toward more transparency on research data.[5]

These broad principles seem cogent to me, and the fashion is unlikely to be transitory, but the implementation is less straightforward. There are another set of issues and valid interests to consider. Those who give their personal data and samples for research have a right to protection of their confidentiality, and to having proper attention paid to any restrictions in the consent they have given, and putting personal data into the public domain could leave research volunteers open to having their anonymity broken. Researchers who have worked long and hard accumulating data have a right to appropriate academic recognition, and sometimes to claims of intellectual property. Ignoring these rights would be unfair and foolish—without incentivized, motivated researchers, and research subjects, there would be no data to share (for example, the Expert Advisory Group on Data Access Report on Incentives to Support Data Access.[6])

Some data can just be put out for anyone to see (as was the human genome sequence), but for clinical studies, quite often data release is "managed" or "controlled," for example, by having an application process overseen by a data access committee, and imposing various conditions on secondary users of the data. Restrictions that potential users might be asked to agree include, for example, undertaking not to pass data on to third parties without permission, not to publish in certain delimited areas for a specified period to allow the primary collectors of the data to publish first, and not to attempt in any way to reidentify or contact anonymized data subjects. Issues of governance and control are still evolving. As referred to, the National Institutes of Health has recently set out quite explicit data sharing guidelines, and work is now under way by Expert Advisory Group on Data Access to recommend appropriate guidelines for governance and resourcing of research data release, for U.K. funding agencies. Other public facing organizations, such as the Global Alliance for Genomics and Health,[7] are working toward better tools and standards for responsible data sharing.

And then there is cost. Releasing large sets of data in a useful way is not free goods—it takes time and effort, and has associated costs. These can often be met from within a study budget, but there is an increasing recognition of the need for planning, costing, and resourcing explicit data release

---

[1] *University of Cambridge, United Kingdom.*

The author declares no funding or conflicts of interest.

Correspondence: Martin Bobrow, FRCP, FMedSci, FRS, Hon Faculty, Wellcome Trust Sanger Institute, Hinxton, United Kingdom. (mb238@cam.ac.uk).

ISSN: 0041-1337/15/9904-654

DOI: 10.1097/TP.0000000000000714

policies from the outset, at the time the work is first initiated and grants applied for.

The complexities are real, and must be dealt with, but they should not be used as excuses for trying to avoid sharing data. Scientists who claim they are protecting their patients, when they are obviously really protecting themselves, do not look good in today's world. The mood has changed, and although the details of how to manage it all are still in flux many researchers are trying to find sensible ways of balancing the risks. Hopefully, those responsible for credit and career advancement in the academic world (and even in industry) will soon show recognition of the fact that having others use your data is a sincere form of flattery, and deserves as much or more credit than having a few papers published.

REFERENCES

1. Hutchinson JA. Data sharing. *Transplantation*. 2015;99(4):649–650.
2. http://www.casimir.org.uk/storyfiles/64.0.summary_of_bermuda_principles.pdf.
3. http://www.wellcome.ac.uk/stellent/groups/corporatesite/@policy_communications/documents/web_document/wtd003207.pdf.
4. http://www.nature.com/nature/journal/v461/n7261/pdf/461168a.pdf.
5. www.alltrials.net.
6. http://www.wellcome.ac.uk/stellent/groups/corporatesite/@msh_peda/documents/web_document/wtp056495.pdf; http://www.knowledge-exchange.info/Default.aspx?ID=733.
7. http://genomicsandhealth.org.