Data sharing

Data sharing is increasingly viewed as an essential step in improving research transparency and reproducibility (Taichman *et al*, 2016; Vickers, 2006). There has been a lot of discussion on the imperative for data sharing in the biomedical arena, particularly of publically funded research. As a result, there are many disciplines where proposals for data sharing are being discussed.

Publishers, including PLOS and the BMJ Publishing Group, that have implemented data sharing requirements have found that it is not trivial to do. A recent post on the Scholarly Kitchen (http://scholarlykitchen.sspnet.org/2016/01/13/what-price-progress-the-costs-of-an-effective-data-publishing-policy/) discussed the costs and workforce issues that could result from making data sharing a requirement for publication of research. Some of the questions raised relate to availability of data archives and infrastructure outside of publishers, and the enforcement policies of journals to assure compliance with data archiving and sharing. If there is no enforcement of a data sharing policy, and the infrastructure to support data sharing is lacking (it is currently patchy, although well developed in Australia, for example) and the editors do not have a policy of peer review of the data prior to acceptance of the paper, how will requiring data sharing actually improve the integrity of the research? Other questions include how long should data remain available, who should manage the availability and sharing of the data, and how much will these requirements cost? In addition, there is a need to ensure that those whose data are reused get adequate credit—something that is not routinely done, but which groups such as Force 11 and others (https://www.force11.org/group/joint-declaration-data-citation-principles-final) are working towards. Most recently, questions about the legitimate requests for and re-use of data have been explored systematically and thoughtfully by Lewandowsky & Bishop (2016).

In light of the above concerns about implementing a data sharing policy, COPE invites discussion on this topic, specifically relating to the following questions:

1) Should researchers be required now to make their data available as a condition of publishing?
2) Who should disseminate guidelines and/or monitor data sharing practices?
3) What issues surround the re-use of published data?
4) Should data deposited by authors be subject to peer review? Prior to publication? To settle disputing claims about results? For use in systematic reviews and meta-analyses?
5) What best practices are already used by journals, publishers and data repositories that could be adapted for use by others considering data sharing requirements?
6) Since past practices do not often enable data sharing in any easy way, should there be an 'amnesty' for old work, but stricter standards applied to work now being done?

References

Lewandowsky S, Bishop D. (2016). Research integrity: Don't let transparency damage science. *Nature*, 25 January, vol 529; http://www.nature.com/news/research-integrity-don-t-let-transparency-damage-science-1.19219

Taichman DB, *et al*. (2016). Sharing clinical trial data—A proposal from the International Committee of Medical Journal Editors. *New England Journal of Medicine,* January 20. DOI: 10.1056/NEJMe1515172

Vickers A J. (2006). Whose data set is it anyway? Sharing raw data from randomized trials. *Trials 7; 15.* DOI: 10.1186/1745-6215-7-15

COMMENTS FROM THE FORUM (Friday 12 February 2016) – NOTE, Comments do not imply formal COPE advice, or consensus.

- What are the practical issues, rather than just the ethical issues, surrounding data sharing? We need to understand the different problems in different research communities. Are there examples of good practices and cultures in certain research communities? Perhaps COPE could highlight minimal standards, with features or examples of good practice? This could range from a 'basic' level (ie, encourage authors to make 'data availability statements' but allow authors to declare data are only available on request) through to enforcing data sharing by specifying sharing in a repository, for example.
- Different academic communities/fields are at different stages in introducing policies on data sharing, and are in different states of readiness to share research data. This is related to whether or not these communities have data repositories, and whether there are good practices and cultures in these research communities
- We need to consider the practical implementation of data sharing policies as a continuum of how journals should encourage, require or enforce data sharing. Some journals, such as PLOS and the BMJ, have a strong data policy. Strong data sharing policies are good for science and research, but do require a lot of resources to do properly. Some journals have a policy where data are required on publication, but this is not easy to implement.
- Minimum standards are necessary, but we need to look at the process much earlier on and engage with funders, institutions and the larger community to set those standards—otherwise we will struggle to get authors to meet them. Publishers and editors cannot do this on their own. A much more global approach is needed.
- Editors need support from funders and the wider community. Funder support is growing, especially as funders are aware that the impact of the research they fund can be maximised if the underlying data are provided**.** There are mandates from some UK funders regarding data sharing.
- Different communities and disciplines have very different attitudes to data sharing. Data sharing is very strong in the biosciences and computer science, for example, but in other disciplines there is less interest. Perhaps we need to search for what is already out there and what works, and then see how we set the standards.
- Different communities have different views on what 'data' are—there are no specific definitions of what the underlying data for a specific paper should be. What are data? For some disciplines/communities it may be more appropriate to share methods/computer code rather than data. And what is the difference between data source and data?
- We also need to consider quality control of the data deposited, and whose responsibility this is. The data also have to be in a form that is accessible and usable for researchers in the future. Should these checks be handled at the repository level rather than the editorial/journal level? If the controls are at the repository level, then journals can require authors to put their data in that repository, and the repositary can stipulate the requirements and format.
- Not all communities have date to share—do mathematicians have data to share for example? Hence this discussion is only relevant for data intensive communities.

- The biggest barrier to data sharing is cultural and getting researchers to behave differently. There are things that publishers could be doing around infrastructural changes before starting to mandate data sharing. What does the infrastructure need to be—data repository, something the publisher has to set up, or something the funders need to be involved in and co-ordinate/reinforce?
- There can be no one size fits all—not all guidelines will be applicable to all disciplines. We may need to distinguish between data related to 'biomedicine' and everything else 'outside biomedicine'. Clinical trial data has forced this issue into the main stream. But a statement of ideals could apply to all disciplines. Any guidance should also include clarity about the granularity issue of both the data (raw data, the analysis dataset which is more refined than the raw data) and the scope issue (should the full data that might exist be made available or just those data related to the current publication)?
- Lack of reproducibility in the basic open sciences has also driven the move towards open data.

ACTION: COPE is currently conducting desk research with attention to ethical issues and interdisciplinary differences in current practice related to data sharing. This discussion will form part of a larger document and possible guidance on this issue for our members.

COMMENTS POSTED ON THE WEBSITE

Posted by Richard Tol, 1/2/2016

Without data and code, replicability is a farce. Hence, data and code should be available as much as possible. Arguments about effort are nonsense. If your data are sufficiently organized for analysis, then it is a small additional effort to upload the files to some repository.

I disagree with Lewandowsky and Bishop. Irony apart -- Lewandowsky has a reputation for hiding sloppy research, Bishop played a small but key role in the harassing of Tim Hunt -- they argue for reduced transparency so as to protect researchers against naughty outsiders. This does not work. If they want to beat you, they will find a stick. Hiding your data just hands them a bigger stick. At the same time, reduced transparency is effective in protecting naughty researchers.

Posted by Paul Matthews , 1/2/2016

"Most recently, questions about the legitimate requests for and re-use of data have been explored systematically and thoughtfully by Lewandowsky & Bishop (2016)."

This is a jaw-dropping and quite worrying remark from the anonymous COPE representative who wrote this piece.

If this is the view of COPE, there is no point in taking part in your Webinar.

Please read the comments under the Nature article, and the following blog posts by four academics:

Nicole Janz

Getting the idea of transparency all wrong

Judith Curry

Violating the norms and ethos of science

James Coyne

Further insights into the war against data sharing: the Science Media Centre's letter writing campaign to UK parliament

and myself
Nature on research integrity?
(links removed as they triggered the sp@m trap)

Posted by Wolff-Michael Roth, 1/2/2016
1. I would distinguish between data source and data---especially important in qualitative research. The data are what is used to support claims made, whereas a video or transcript is a data source, and different pieces may be extracted to support very different claims.
2. Multiple uses of large data sets are already common in the statistical analysis of PISA and other social science data. In any case, use of such data sets, at least in Canada, has to be approved by the local Research Ethics Board.
3. Provisions need to be made during recruitment and informed consent that data, once published, can no longer be withdrawn, as is current practice ("may withdraw at any time")

Posted by William Grant, 1/2/2016
Three issues; First, what constitutes a 'clinical trial" - - only publicly funded trials? Many residents and fellows undertake trials which could be construed as clinical (comparison groups, interventions, specific outcome measures). These trials are often small in number making potential patient identification more problematic. Not certain that these trials could even be blinded sufficiently to protect patients.
Second, who 'owns' the data? What prevents another individual from making a career wading through these public data sets publishing like crazy using the data inappropriately in manners in which the data was never intended?
Third, what it the time limit for sharing data. Some data will clearly have a shelf life due to advancements in the field and will become out-of-date or otherwise inappropriate. What prevents someone from using older data?

Posted by Charon Pierson, 1/2/2016
I appreciate the comment of Wolff-Michael Roth about the distinction between data source and data but I think there remains a lack of confidence on the part of many in the data source for some qualitative research. For example, Carlos Castaneda received his PhD in Anthropology from UCLA based on "field notes" from his experiences with shamanism and peyote and went on to write numerous "non-fiction" books about the Yaqui Shaman Don Juan Matus (https://en.wikipedia.org/wiki/Carlos_Castaneda). Critics largely claim Castaneda's work was fiction not an ethnography, and although his PhD was never revoked, a former Chair of Yale's Anthropology Department claims: "to me it remains a disturbing and unforgivable breach of ethics" (http://www.salon.com/2007/04/12/castaneda/). Perhaps this would not occur in today's academic climate, and perhaps a more rigorous peer review would have noted discrepancies in Castaneda's work, but the data and the data source can be difficult if not impossible to disambiguate.

Posted by Aliaksandr Birukou, 12/2/2016
From the computer science perspective, I'd like to add:
1. Publishing software should be separated from publishing data. Here is some info on experimenting with sharing software and checking it during review:http://cacm.acm.org/magazines/2015/3/183593-the-real-software-crisis/

2. When we move to other disciplines (e.g., computer science), formal peer-reviewed conference proceedings play a very important role, maybe even bigger role than many journals. Thus, data sharing should be also supported for conferences and in proceedings, not only for journals.
Aliaksandr Birukou

Posted by Ian Blackford, 15/2/2016
I'm aware the date of the forum was last Friday, but I hope it is still OK to comment.
My interest is whether journals should require reviewers to review the deposited data. As we already ask so much of reviewers I fear asking them to review large datasets will take up more of their time and perhaps make it even harder for Editors to find willing reviewers. I'm certainly not against sharing data, I'm just trying to get an understanding of how others feel data should be reviewed.