Title: Dryad in the UK and USA - prospective and retrospective data publication

Authors: Kursheed Khan, Andrew Weeks

Kursheed Khan, Medical Student, University of Liverpool, Liverpool, UK

Andrew Weeks, Professor of International Maternal Health, Department of Women's and Children's Health, University of Liverpool, Liverpool, UK

Competing interests: No competing interests.

All authors have completed the ICMJE uniform disclosure form at www.icmje.org/coi_disclosure.pdf and declare: no support from any organisation for the submitted work; no financial relationships with any organisations that might have an interest in the submitted work in the previous three years; no other relationships or activities that could appear to have influenced the submitted work.

Dryad in the UK and USA - prospective and retrospective data publication

Health-related research in the UK is estimated to cost around £8.5 billion per annum and the National Institute of Health invested $32.3 billion in the US [1, 2]. This research, especially clinical trials, generates a huge amount of data, much of which carries importance far beyond the primary analysis. Although the trial team may publish secondary analyses, it is not uncommon for the study team to move on to the next project without fully exploiting the database. This costly and hard-won data may then sit on a computer for many years before finally being discarded.

In recent years there has been a drive to publish clinical datasets, as well as reporting findings. It is most common to do this at the same time or shortly after the paper has been published. Dryad is a digital repository recommended by Toxicological Sciences which stores data from publications for free access. At Toxicological Sciences, data sharing is voluntary and highly encouraged, and it is free for authors to do so as the costs are covered by the journal.

Prospective data publishing refers to publishing datasets alongside the original paper with the primary analysis. Retrospective data publishing is when datasets are publishing after the research paper is published.

Based on our recent experience of publishing a retrospective dataset [3, 4], this article aims to provide authors with a guide on how to prospectively or retrospectively publish their data. The data is citeable and freely accessible to all web users.

Importance of data sharing

Making the trial dataset publicly available allows other study teams to explore their own hypotheses, to conduct individual patient data meta-analysis, and obtain data in their preferred form. It also allows the primary analysis to be checked. According to the Wellcome Trust, transparency should be encouraged as it leads to higher quality research, better value for money and higher quality science. [5] This is important not only for unpublished trials, but also for those that are already published.

Whilst the sharing of data has multiple benefits, there are also dangers of not sharing data. Incomplete analysis of data reduces the amount of information about the condition under study. Ultimately, this can compromise treatment choices leading to lower levels of health and patient care. [6] It can also allow errors in the analysis to go unrecognised leading to the publication of inaccurate results and conclusions. Most commonly this is accidental, but there have been repeated episodes where results have been fabricated, and the risk of this is also reduced with data publication. In the long run data sharing leads to a higher quality of research, with subsequent individual benefits through improved patient care.

Some academics have concerns about releasing data, fearing that they will lose control of their project and data. It is feared that others may rush to conduct inappropriate secondary

analyses, or that those with vested commercial or academic interests may seek to misuse the data to reach contradictory conclusions. If there are these fears are justified then it is reasonable to state a time at which the data will be published (e.g. 6 months or 2 years after the study ends), thus allowing the primary research team to complete all their secondary analyses first.

Anonymising data

Before any data is submitted, measures must be taken to ensure the data is anonymised to an acceptable level. This is especially important with open access data repositories like Dryad where anyone can view the information.

Hrynaszkiewicz et al, suggest minimum standards to ensure confidentiality with data sharing.[7] An important distinction is made between direct and indirect patient identifiers. Direct identifiers, such as name or date of birth, have a high probability of being able to identify an individual and so researchers should avoid publishing these if at all possible. In contrast, indirect indicators such as area of residence or hospital site have a low risk of accurately identifying an individual, even with multiple indirect variables. The authors recommend using less than 3 indirect identifiers when publishing data. More indirect (or any direct) identifiers may compromise anonymity and patient confidentiality. If three or more indirect identifiers (or any direct identifiers) are used, explicit justification must be given and permission sought from an independent ethics committee. [7]

There are multiple formats in which to store the data, but it is suggested that databases are stored using Microsoft Excel. [7] Although it is not as comprehensive as specialist databases such as SPSS, it is widely used and so makes the data more accessible. It also suggested data must be "cleaned", removing errors, missing data and duplicate information. The data should also be "well annotated", meaning that any coded headings or short abbreviations must be fully explained. This is particularly a problem with Excel where the headings are usually shortened (e.g. 'OXYPRE' might be used as an abbreviation for 'oxytocin infusion used prior to birth for augmentation of labour') and data within databases is commonly coded (e.g. 1 for 'yes', 2 for 'no' and 999 for 'missing data'). If this is the case the codes may need to be lengthened, or a supplementary page of descriptions and definitions may need to be added. Without the decoding, the data becomes difficult or impossible to interpret, and this can itself lead to errors in analysis and interpretation. [8]

Consent for data publication should be sought directly from the study participants if possible. If explicit consent has not been gained, then appropriate reasons must be given. This will be no problem for those studies that have planned to share the data from the start. For those seeking to share data retrospectively, however, obtaining informed consent from each patient is very unlikely to be achievable. In this case it is suggested that permission be sought from the local Caldicott Guardian, as well as the ethics committee who originally approved the study. They would usually be happy for this so long as the appropriate measures have been taken to anonymise the data.

For those publishing data retrospectively, consent for publication also needs to be sought from all authors of the original publication, as the data needs to be released under a Creative Commons 0 license (CC0). A CC0 license is a 'no rights reserved' licence and waives the restriction of copyright law. This means the data may be used freely and without restriction for future user, so that they may enhance and build upon the data. [9]

Data submission

Whether data is being submitted retrospectively or prospectively, the submission procedure is the same. Dryad accepts data in all forms so that any research team can submit datasets to the repository. Each dataset is given a Direct Object Identifiers (DOI), which can be used to cite the dataset and to access it. This gives authors credit for use of their datasets.

If submission is prospective the dataset can be referenced within the published paper, with the DOI number. Dryad offers a the option of keeping the data private during peer review and then making it public once the paper with the primary analysis has been published.

If the dataset is published retrospectively, the process remains the same - the only difference being that the DOI cannot be easily added or linked to the original publication. There are ways around this. For our retrospective data publication, the link was made in a BMJ letter, [4] whilst for those journals with online response systems (such as the BMJ), the DOI can be submitted as a rapid response to the article. This can help future users to access the dataset.

Dryad states on their website that the submission procedure usually takes less than 15 minutes. Therefore it is a relatively easy process once the dataset has been prepared.

The Future

In this article so far we have only discussed the possibility of publishing clinical trial data, however even the most basic scientific experiments produce data. With online repositories such as Dryad, it is now possible to have all raw datasets published relatively easily. Dryad guarantees data for 10 years to be accessible with no extra cost to the author.

This would mean datasets from large scale trials to basic scientific papers will be available to access. The raw data can be compiled and even used to create new analyses without having to necessarily preform any experiments. With the help of Dryad, authors of the original dataset can be credited for their contributions. Therefore this can be of benefit to the research community as a whole.

The field of toxicology is always looking to improve the quality of scientific research. At Toxicological Sciences it is not only the accuracy of results which is emphasised, it also the reproducibility of the results. [10] Research in toxicology is not primarily focussed on new treatments but rather preventing harmful effects. This is of particular significance, as hypothetically data can be analysed from multiple sources on the same treatment and correlation of side effects or potential risk factors can be extrapolated.

At Toxicological sciences it is encourage to share data not to treat it as a protected possession but rather something openly shared, and with the correct systems in place this is possible.[11]

Conclusion

Data publication is becoming an increasingly important part of research dissemination - indeed some major funders now require it as a condition of their funding. The process is simple and we would encourage those with datasets from previous pieces of research to increase its value by sharing their data in the way in which we have outlined.

Fig 1 – how to submit data to Dryad[12]

References

1) UK Health Research Classification System. UK Health Research Analysis 2014 http://www.hrcsonline.net/sites/default/files/UKCRCHealthResearchAnalysis2014%20WEB.pdf (Accessed 16/11/15)
2) National Institute of Health. Budget Note. http://www.nih.gov/about-nih/what-we-do/budget#note. (Accessed 20/04/16)
3) Dryad. Data from: Umbilical vein oxytocin for the treatment of retained placenta (Release Study): a double-blind, randomised controlled trial. http://datadryad.org/resource/doi:10.5061/dryad.g3gj1.(Accessed 9/04/15)
4) Khan K, Weeks AD. Example of retrospective dataset publication through Dryad. BMJ 2015;350:h1788
5) Wellcome Trust. Sharing research data to improve public health: full joint statement by funders of health research. http://www.wellcome.ac.uk/About-us/Policy/Spotlight-issues/Data-sharing/Public-health-and-epidemiology/WTDV030690.htm. (Accessed 20/04/16)
6) Lehman R, Loder E. Missing clinical trial data. BMJ 2012;344:d8158
7) Hrynaszkiewicz I, Norton ML, Vickers AJ, Altman DG. Preparing raw clinical data for publication: guidance for journal editors, authors, and peer reviewers. BMJ. 2010 Jan 28;340:c181.
8) White EP, Baldridge E, Brym ZT, Locey KJ, McGlinn DJ, Supp SR (2013) Nine simple ways to make it easier to (re)use your data. Ideas in Ecology & Evolution 6(2):1–10. http://doi.org/10.4033/iee.2013.6b.6.f(Accessed 15/10/15)
9) Creative commons. About CC0 — "No Rights Reserved".https://creativecommons.org/about/cc0 (Accessed 15/10/15)
10) Miller, G. W. (2014). Improving Reproducibility in toxicology. Toxicol. Sci. 139, 1–3.
11) Miller, G. W. (2015). Data sharing in toxicology: beyond show and tell. Toxicol. Sci. 143, 3–5.
12) Dryad. Frequently asked questions. http://datadryad.org/pages/faq.(Accessed 15/10/15)