Dryad in the UK and USA—Prospective and Retrospective Data Publication

Kursheed Khan* and Andrew Weeks†,1

*University of Liverpool, Liverpool, UK and †Department of Women’s and Children’s Health, University of Liverpool, Liverpool, UK

1To whom correspondence should be addressed at E-mail: hlkkhan@liv.ac.uk

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 (National Institute of Health, 2016; UK Health Research Classification System, 2014). 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 (Dryad, 2015; Khan and Weeks, 2015), this article aims to provide authors with a guide on how to prospectively or retrospectively publish their data. The data is citable 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 (Wellcome Trust, 2016). This is important not only for unpublished trials, but also for those that are already published.

Whereas 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 (Lehman and Loder, 2012). It can also allow errors in the analysis to go unrecognized 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 (eg, 6 months or 2 years after the study ends), thus allowing the primary research team to complete all their secondary analyses first.

ANONYMIZING DATA

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

Hrynaszkiewicz et al. (2010) suggest minimum standards to ensure confidentiality with data sharing. 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 3 or more indirect identifiers (or any direct identifiers) are used, explicit justification must be given and permission sought from an independent ethics committee (Hrynaszkiewicz et al., 2010).

There are multiple formats in which to store the data, but it is suggested that databases are stored using Microsoft Excel (Hrynaszkiewicz et al., 2010). 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 (eg, “OXYPRE” might be used as an abbreviation for “oxytocin infusion used prior to birth for augmentation of labor”) and data within databases is commonly coded (eg, 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 (White et al., 2013).

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 anonymize 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” license 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 (Creative commons, 2015).

**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 (Figure 1).

If submission is prospective the dataset can be referenced within the published paper, with the DOI number. Dryad offers 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 (Khan and Weeks, 2015), whereas
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 perform 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 emphasized, but it also the reproducibility of the results (Miller, 2014). Research in toxicology is not only primarily focused on new treatments but also rather preventing harmful effects. This is of particular significance, as hypothetically data can be analyzed 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 (Miller, 2015).

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.

REFERENCES
Creative commons. (2015). About CC0—“No Rights Reserved”. Available at: https://creativecommons.org/about/cc0. Accessed October 15, 2015.
Wellcome Trust. (2016). Sharing research data to improve public health: full joint statement by funders of health research. Available at: http://www.wellcome.ac.uk/About-us/Policy/