The path to open data

The increasing volumes of biological and clinical data have the potential to greatly enhance our understanding of the processes underlying kidney function and disease. However, maximizing outputs from these data requires a collaborative and open approach to data sharing that can only be achieved through united efforts by researchers, funders and publishers.

The growth of data in the medical sciences has been fuelled by technological advances and by reductions in the cost of high-throughput technologies. In nephrology, these developments have led to unprecedented insights into physiological and pathophysiological processes. They have enabled increased understanding of the complexity of kidney cell types and the contribution of genetic variants to rare and complex diseases, and led to the identification of potential therapeutic targets. Data sets are powerful resources that should be optimally mined to ensure maximum output. In this issue of *Nature Reviews Nephrology*, Sampson and Kang[1] argue that the optimal use of data sets is best accomplished by facilitating and embracing widespread data sharing.

Data sharing increases sample sizes, enables comparisons between different populations and facilitates replication analyses. These benefits extend beyond 'big data' analyses from omics studies. The International Committee of Medical Journal Editors (ICMJE) believes that researchers have an ethical obligation to responsibly share data generated by interventional clinical trials given the risks associated with participating in such studies[2] — a view that seems to be shared by the majority of trial participants[3]. In instances where lack of patient consent does prevent data sharing, it would be helpful to understand whether consent was denied or not requested by publicly sharing the consent form used. Beyond consent, other barriers — all of which can be overcome — still exist.

The curation of data in a meaningful way is time and resource consuming — planning and expertise are required to ensure that data are FAIR (that is, Findable, Accessible, Interoperable and Reusable by both humans and machines)[4]. Standardization of protocols for data generation and analysis is an important first step to enable sharing and re-use of data. In their comment, Sampson and Kang[1] highlight the importance of careful study design and the need to establish collaborative data resources that focus on kidney disease. The enrichment of data repositories with kidney-relevant data is indeed needed, but although disease-specific resources could provide strong curation and enforce data standards, arguably this enrichment should occur in central resources that are supported by the wider community and might be more stable in the long-term.

In a competitive research environment, the move to an open data approach will only be successful if efforts to enable and promote sharing are recognized and rewarded. Researchers who create robust data sets should receive credit for their contribution to the community, for example, through the assignment of unique persistent identifiers (PIDs) to data sets[5]. Citation of PIDs provides recognition that data sets are a valuable scientific output and is encouraged by the *Nature Research* journals[5,6]. The tracking of publications linked to a PID enables assessment of the reach and impact of the data set. Such contributions must be recognized by funding bodies and tenure committees, as rewarding the generation of robust data sets that stand up to scrutiny in terms of their quality and annotation is vital to the success of an open data approach. Publishers also have a role here. ICMJE requires that published clinical trials contain a data sharing statement, but does not mandate data sharing[2]. All *Nature Research* journals publish data availability statements and express a preference for data sets to be shared via public repositories, but this is only mandated for certain data sets — a stance that could be strengthened in the future.

The expertise of researchers who contribute to scientific discovery through independent analyses of open data sets should also be recognized. These researchers must appropriately credit the curators of the original data. Publishers must assess the robustness of secondary analyses through peer review and mandate that all data sources are acknowledged in a trackable manner.

The research community is increasingly embracing data sharing but more work is needed to overcome the remaining barriers. Funders, institutes and publishers should promote progress in this area by providing support and appropriate recognition to the researchers who create and use open data sets.

1. Sampson, M. G. & Kang, H. M. Using and producing publicly available genomic data to accelerate discovery in nephrology. *Nat. Rev. Nephrol.* https://doi.org/10.1038/s41581-019-0166-z (2019).
2. Taichman, D. B. et al. Data sharing statements for clinical trials: a requirement of the International Committee of Medical Journal Editors. *JAMA* **317**, 2491–2491 (2017).
3. Mello, M. M. et al. Clinical trial participants' views of the risks and benefits of data sharing. *N. Engl. J. Med.* **378**, 2202–2211 (2018).
4. Wilkinson, M. D. et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **3**, 160018 (2016).
5. Pierce, H. H. et al. Credit data generators for data reuse. *Nature* **570**, 30–32 (2019).
6. Data citation needed [Editorial]. *Sci. Data* **6**, 27 (2019).