International Journal of Epidemiology, 2017, 1729–1733
doi: 10.1093/ije/dyx195
Advance Access Publication Date: 15 September 2017
Editorial

Editorial

Policies and strategies to facilitate secondary use of research data in the health sciences

Paul R Burton,[1†]* Natalie Banner,[2†] Mark J Elliot,[3] Bartha Maria Knoppers[4] and James Banks[5]

[1]Institute of Health and Society, Newcastle University, Newcastle upon Tyne, UK, [2]Wellcome Trust, Headquarters, 215 Euston Road, London, UK, [3]School of Social Sciences, University of Manchester, Manchester, UK, [4]Centre of Genomics and Policy, McGill University, Montreal, QC, Canada and [5]Institute for Fiscal Studies and University of Manchester, Manchester, UK

*Corresponding author. Institute of Health and Society, Newcastle University, Baddiley-Clark Building, Newcastle upon Tyne NE2 4AX, UK. E-mail: paul.burton@newcastle.ac.uk
[†]Joint first authors

Data are increasingly seen as a fundamental resource that underpins research across biomedicine and the broader health sciences. They often have important value and utility beyond the purpose for which they were originally collected. Recognizing this, many initiatives across the globe actively seek to enable and promote greater sharing of research data, and most major funders now require researchers to set out formal plans for managing and sharing their data with users beyond their own study team. To provide a contextual backdrop to the article and to point the reader to additional sources of relevant information, we include a table listing examples of international, regional and national projects that seek to enhance and facilitate data sharing in a variety of different ways. Table 1 is not intended as an exhaustive catalogue, rather it aims to provide an illustrative listing of a variety of projects we believe to be in good standing around the world, to demonstrate the broad diversity of responses to the important challenges presented by data sharing.

Despite the acknowledged and growing importance of data sharing, individual researchers and research groups rightly perceive that the primary pressures on them when applying for grant funding are to demonstrate scientific rigour and the potential to undertake and publish innovative research. As a result, the practicalities and governance of data management—including data sharing and the provision of secondary data access—are often seen as being subordinate; the need to produce a data management/sharing plan is widely viewed as a bureaucratic inconvenience with limited relevance to the actual science of study design. However, a similar attitude used to apply to the formal designation of criteria for identifying and assessing papers before undertaking a systematic review or meta-analysis, and the reporting of, and realistic power calculations for, genetic association studies. These same tasks are now seen as being central to good scientific design,[1–3] and reviewers are expected to comment upon them from a scientific perspective. Similarly, researchers proposing a study involving health-related data on individual human subjects should now expect to have to provide a comprehensive data management plan ensuring that study data will be exploitable in all ways needed to achieve stated scientific objectives, including making data available for secondary users where that is appropriate. Such plans should be realistic, consistent with all relevant governance documentation, attuned to the contemporary regulatory and social landscapes and robust to likely future changes as those landscapes evolve. Furthermore, it should be expected that the data management plan will rigorously be assessed as a core element of grant review.

Given the importance of these issues, funders and international experts are increasingly focused on supporting and promoting better governance of data and better understanding of that governance. In particular, it is recognized that in the absence of such an effort there is a significant risk that overly strict or inappropriately lax governance

Table 1. Examples of international initiatives contributing to data sharing

Example	Purpose
Initiatives spanning genomics and phenotypic data-sharing	
NIH BD2K programme https://datascience.nih.gov/bd2k/about	BD2K (Big Data to Knowledge) is a US (trans-NIH) initiative established to enable biomedical research as a digital research enterprise, to facilitate discovery and support new knowledge and to maximize community engagement. One major aim is to facilitate broad use of biomedical digital assets by making them Findable, Accessible, Interoperable and Reusable (FAIR)
ELIXIR project https://www.elixir-europe.org/	ELIXIR unites Europe's leading life science organizations in managing and safeguarding the increasing volume of data being generated by publicly funded research. It coordinates, integrates and sustains bioinformatics resources across its member states, and enables users in academia and industry to access services that are vital for their research
BBMRI (Biobanking and BioMolecular Resources research Infrastructure) http://www.bbmri-eric.eu/BBMRI-ERIC/about-us/	BBMRI has developed, established and operates a pan-European distributed research infrastructure of biobanks and biomolecular resources in order to facilitate the access to resources as well as facilities, and to support high quality biomolecular and medical research
RDA (Research Data Alliance) https://rd-alliance.org/	The Research Data Alliance (RDA) was created as a community-driven organization in 2013. It was launched by the European Commission, the United States National Science Foundation and National Institute of Standards and Technology, and the Australian Government's Department of Innovation, with the goal of building the social and technical infrastructure to enable open sharing of data
Initiatives emerging from the genomics community	
GA4GH (Global Alliance for Genomics and Health) http://genomicsandhealth.org/	Accelerating progress in human health by helping to establish a common framework of harmonized approaches to enable effective and responsible sharing of genomic and clinical data, and by catalyzing data-sharing projects that drive and demonstrate the value of data-sharingGA4GH Catalogue of Global Activities eHealth (currently 85 listed):http://genomicsandhealth.org/work-products-demonstration-projects/catalogue-activities-ehealth-0
ICGC (International Cancer Genome Consortium) http://icgc.org/daco	Generating comprehensive catalogues of genomic abnormalities (somatic mutations, abnormal expression of genes, epigenetic modifications) in tumours from 50 different cancer types and/or subtypes which are of clinical and societal importance across the globe, and making the data available to the entire research community as rapidly as possible, and with minimal restrictions, to accelerate research into the causes and control of cancer
EGA (European Genome-phenome Archive) https://www.ebi.ac.uk/ega/about/introduction	The EGA provides a service for the permanent archiving and distribution of personally identifiable genetic and phenotypic data resulting from biomedical research projects. Data at EGA were collected from individuals whose consent agreements authorize data release only for specific research use to bona fide researchers. Strict protocols govern how information is managed, stored and distributed by the EGA project
dbGaP (Database of Genotypes and Phenotypes) https://www.ncbi.nlm.nih.gov/gap	The database of Genotypes and Phenotypes (dbGaP) was developed by NIH to archive and distribute the data and results from studies that have investigated the interaction of genotype and phenotype in humans
Café Variome http://www.cafevariome.org/about/cafevariome	Cafe Variome is not a database but a 'shop window' for what exists in various data sources. It is designed to enable users to ask the question 'Where can certain data be found?' Once that question is answered, access to those data may be sought under one of the three different models as stipulated by the data owner/submitter
Clinical research data: clinical trials and health service data	
YODA (Yale Open Data Access project) http://yoda.yale.edu/welcome-yoda-project	Advocating the responsible sharing of clinical research data, particularly clinical trials data, with a commitment to open science and data transparency. The mission of the YODA project is to not only increase access to clinical research data, but also to promote its use to generate new knowledge.
CSDR (ClinicalStudyDataRequest.com) http://www.clinicalstudydatarequest.com	A consortium of 13 (currently) pharmaceutical companies providing controlled access to data from clinical research studies, particularly clinical trials

(Continued)

Table 1. Continued

Example	Purpose
Farr Institute http://www.farrinstitute.org/	The Farr Institute does not own or control data but works across diverse domains to enhance the value and useability of health-related data (infrastructure and tools, regulation and ethics, capacity building, public engagement, professional partnerships). Its aim is to enable the application of cutting-edge data science to address major challenges across the nation's 65 million population
CHC (Connected Health Cities) https://www.connectedhealthcities.org/	There are four Connected Health Cities (CHCs) in the North of England. Each local city region aims to unite health and social care services so that together they can share information and improve the health of local people. The efficient use of health technology and data lies at the heart of CHC
Initiatives focusing on interoperability	
P^3G's IPAC (The Public Population Project in Genomics and Society's International Policy interoperability and data Access Clearinghouse) http://www.p3g.org/ipac	Assists researchers to understand, work with and meet diverse ethical and legal regulatory requirements both nationally and internationally
Maelstrom Research https://www.maelstrom-research.org/	Maelstrom Research offers a suite of methods and software tools, as well as expertise, to partners to facilitate data documentation, harmonization and integration. The Maelstrom Repository [https://www.maelstrom-research.org/repository] is a standardized catalogue encompassing comprehensive information about epidemiological research networks (currently 12 consortia catalogued), epidemiological studies (currently 177 catalogued), the data they collect and their potential for harmonization
METADAC (Managing Ethico-social, Technical and Administrative issues in Data Access) http://www.metadac.ac.uk/metadac-about/	A multi-agency (MRC, ESRC, Wellcome Trust) multi-study data access committee servicing several of the UK's major cohort studies (1958BC, 1970BC, Millennium BC, ELSA, Understanding Society). By identifying and building on opportunities for interoperability and effective professional networking, METADAC aims to provide a scaleable mechanism to incorporate additional cohorts in the future

might be applied in some settings, and this could have serious consequences for scientific utility, validity, the public good and the philosophy of open science. The Expert Advisory Group on Data Access (EAGDA), which is convened by four major UK research funders [https://wellcome.ac.uk/what-we-do/our-work/expert-advisory-group-data-access], was therefore asked to explore these issues, and produced a report on the Governance of Data Access.[4] The report highlighted the variability of governance models across different scientific domains, emphasizing the extent to which the whole research community is gradually feeling its way forward in relation to an important topic which we must get right. The full recommendations of the report are set out in Box 1. Although the report was primarily aimed at funders, several key recommendations are directly relevant to the broader research community. This brief commentary provides an overview of these, with particular focus on those recommendations that may be viewed as unexpected or even controversial. The intent is to stimulate discussion across the health science

Box 1. EAGDA recommendations on the governance of data access

1. All project proposals should include data sharing and management plans in funding applications.
2. Funders should support the development of data and metadata standards.
3. Data access processes should be discoverable and transparent for potential data users.
4. Studies should establish proportionate governance mechanisms for data access.
5. Collaboration should not be a necessary condition for data access.
6. Consent should, as far as possible, include provision for further data use beyond the original study.
7. Clear policies should be developed on how depletable resources will be managed.
8. Funders should establish clear penalties and sanctions for breaches of data-sharing rules.
9. Principles of data access should be harmonized as far as possible across studies. Study leaders should also consider whether harmonization of processes is appropriate.
10. Funders should seek to establish the short- and long-term costs of data access, and work to determine when cost-recovery is an appropriate model for studies.
11. Funders should jointly consider how best to sustainably support data-sharing infrastructures.

community. We aim to provoke researchers to think about how they ought best to approach the governance of data management, including the facilitation of data sharing within consortium-based studies and the provision of secondary access to research data beyond the time frame and scope of the main study.

For researchers in the early stages of planning a study, recommendations 6 and 7 highlight two important issues to consider when anticipating potential future uses of study data. First, participant consents should be formulated to facilitate reasonable reuse of the data as far as is possible. There may be good reasons to restrict access for certain specific users or purposes. However, it is not always possible to anticipate the ways in which data might reasonably be used in the future; we therefore recommend that the current default position should be to seek consent that is as broad as possible, provided it does not encompass data uses that would be viewed as controversial or undesirable. Second, it is often appropriate for depletable resources, such as biosamples or tissue specimens, to be managed differently from unrestricted resources that include data or, for example, DNA supplies that may be regenerated from lymphoblastic cell lines. Study leads should consider early on how they wish to manage these differing commodities, ensuring that access processes and protocols are consistent while reflecting the need for policy differences where appropriate. As a specific example, the UK's METADAC (see Table 1) requires full scientific review of any application for a depletable resource—because any suboptimal use represents an opportunity cost for later applications—while avoiding such a review of applications for unrestricted resources.

As research planning moves into the grant writing stage, recommendations 1 to 3 emphasize the importance of developing clear plans for enabling data to be held safely and securely while making them discoverable and accessible to potential users. Discoverability may require data to be deposited in a subject repository or signposted from a portal, both of which necessitate early thinking about data and metadata formats, standards and vocabularies. Funders increasingly require data management plans and should therefore allow the costs associated with implementing them. A robust data management, discovery and access plan can therefore be adequately resourced, but only if it is clearly thought through at the grant-writing stage.

Once a basic data management plan has been constructed, the development of appropriate processes and procedures to ensure sound information governance becomes pivotal, as these are essential if the data are actually to be used and shared. This is the focus of recommendations 4 and 5. Data governance or information governance broadly concerns issues such as data security, ensuring compliance with participant consent, oversight structures, access policies, transparency of processes and management of resources, and appropriate recognition of the professional contribution of a dataset's creators and maintainers to its ongoing use.

Proportionate governance is appropriately calibrated to realistic risk and the resource requirements of making data available. It recognizes that different tiers of access may be appropriate,[5] depending on: the type and sensitivity of data collected; whether it is a depletable resource; the potential it has for reuse in different fields; the terms of consent; and the resources available to support data access. The original custodians have responsibility for the data and, where external data sharing is expected, they must make reasonable efforts to enable others to find and use it. Proportionate governance therefore treads a difficult line. If the level of governance is set incorrectly, there is either a failure to adequately protect data and participants' interests, or excessive restriction on the data's use, through unnecessarily burdensome discovery and access procedures or unwillingness to allow legitimate access.

Recommendation 5 states that collaboration should not be the sole means through which studies seek to fulfil their data-sharing obligations. This suggestion signals a shift away from the default approach adopted by many researchers and research groups. We recognize that it is controversial, and that some researchers feel that opening up access to their research data to secondary users with no previous relationship with the study exposes them or their work to the risk of exploitation and/or suboptimal analysis and interpretation.

There are often very good reasons to collaborate, as data generators and custodians bring substantial expertise and understanding of the research methodologies originally used to collect the data and the subsequent interpretation of those data. It is therefore common practice for data to be shared collaboratively among members of a research group or consortium on a *quid pro quo* basis, and/or for access to data to be granted as part of an agreement to collaborate on analysis and subsequent publications. This can be entirely appropriate, and the EAGDA report does not recommend that these practices should be discouraged. However, the rationale for allowing access to data beyond their original purpose includes enabling novel questions to be asked, different methodologies to be applied, new approaches to the data to be tested and analytical findings to be replicated. All of these will sometimes need to be undertaken independently of the original study team and/or beyond established networks and collaborations. Access policies that demand that data can only be shared with a collaborator, frustrate these benefits and may be perceived by data users and funders to unfairly

limit access. The prevention of findings from being rigorously scrutinized and verified by others runs counter to the ideals of scientific enquiry. Furthermore, the widely used argument that 'only we know the data well enough' might suggest that data generators and funders should have created better metadata and/or provided better support resources to assist qualified others in interpreting the data correctly and meaningfully. Data collectors must be appropriately credited for all uses of their datasets, but collaboration or co-authorship on publications should not be a default requirement for permitting access to data.

Recommendation 9 promotes harmonization of the basic principles underpinning data access, while recognising that the specific processes for data access and management may well need to differ across studies and domains. Harmonization of basic principles has several advantages. It overcomes the 'cottage industry' effect for data users who may currently have to go through multiple similar but idiosyncratic application processes, each addressing different issues; it minimizes duplication of effort by enabling good practice to be shared; and it ensures that data producers, managers and users in a field can all know what to reasonably expect from one another. Harmonization reflects the spirit of data sharing and open science, promoting a collaborative approach to the scientific endeavour.

Conclusions

There is no doubt that data sharing and provision of secondary data access can have a profoundly beneficial impact on progress in biomedicine and the health sciences. However, their management must be thought through very carefully. Study participants, study investigators, research users, funders and society at large all have a stake in ensuring that it is done effectively, proportionally and transparently. At present, the broader scientific and professional research communities are gradually feeling their way forward on a series of issues that have important implications for a broader society that is itself rapidly rethinking many of the key issues that underpin the use and misuse of data. As research data become more discoverable, accessible and useable, it will become easier to learn and understand the ways in which different studies and different fields approach data governance, access and management. This may in itself assist with harmonizing principles of data governance and access. However, sustained focus and coordination from funders, publishers, repositories, institutions, data custodians and users are essential, and the endeavour must be international in scope. It is imperative that better access to data is encouraged while ensuring that information governance remains adequately rigorous, and without creating undue burdens on researchers or undermining the enormous value of the work done by primary data collectors/generators in the first place. There is much ongoing work in this area to develop proportionate and transparent models of access[5] and to establish best practice in different fields.[6,7] We urge strategic thinking across research communities to debate and establish appropriate data governance principles and standards for different settings, and we hope that the publishing of this article in the *International Journal of Epidemiology* will help to promote discussion among an important group of scientific stakeholders.

Acknowledgements

The authors are grateful to Paul Flicek and John Hobcraft for helpful comments on drafts of this editorial, and to funder representatives Jamie Enoch (CRUK), Rebecca Fairbairn (ESRC) and Jon Fistein (MRC) for their input. This article was written on behalf of the Expert Advisory Group on Data Access, drawing on a report produced by the group under the chairmanship of Martin Bobrow (2012–15) and James Banks (2015–).

References

1. Little J, Higgins JP, Ioannidis JP *et al*. Strengthening the reporting of genetic association studies (STREGA): an extension of the strengthening the reporting of observational studies in epidemiology (STROBE) statement. *J Clin Epidemiol* 2009;**62**:597–608 e4.
2. Burton PR, Hansell AL, Fortier I *et al*. Size matters: just how big is BIG?:Quantifying realistic sample size requirements for human genome epidemiology. *Int J Epidemiol* 2009;**38**:263–73.
3. Moher D, Shamseer L, Clarke M *et al*. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Syst Rev* 2015;**4**:1.
4. Expert Advisory Group on Data Access (CRUK E, MRC, WT).*Governance of Data Access*. London: Wellcome Trust, 2016.
5. Burton PR, Murtagh MJ, Boyd A *et al*. Data safe havens in health research and healthcare. *Bioinformatics* 2015;**31**:3241–48.
6. Vasiliki R, Dyke SO, Knoppers BM. An international framework for data sharing: moving forward with the Global Alliance for Genomics and Health. *Biopreservation and Biobanking* 2016;**14**: 256–59.
7. Bobrow M. *How to share your research data and why you should*. 2015. https://wellcome.ac.uk/sites/default/files/how-to-share-your-research-data-eagda-nov15.pdf (5 September 2017, date last accessed).