GOVERNANCE OF DATA ACCESS

June 2015

CANCER RESEARCH UK
E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
MRC Medical Research Council
wellcome trust

EXECUTIVE SUMMARY

Whilst it is widely accepted that data should be shared for secondary research uses where these can be balanced with the maintenance of participant privacy, there has thus far been little oversight or coordination of policies, resources and infrastructure across research funders to ensure data access is efficient, effective and proportionate. Funders have different policies and processes for maximizing the value from the datasets generated by their researchers, and EAGDA wished to identify whether and how improvements to governance for data access and the dissemination of good practice across the funders could enhance this value.

We identified several good practice exemplars for funders and the research community, but also evidence of significant variability between studies that, in practice, lead to inefficiencies in access to and use of data for secondary purposes. We suggest that an approach that is more transparent and coordinated among research funders would provide benefits, both in terms of increasing the efficiency of data access and in ensuring that good governance approaches can get on with the job of ensuring the right balance between protecting research participants' interests and enabling access to data for further research.

This report makes several recommendations to funders, many of which are interlinked. These focus on areas in which specific funder action could improve the governance of data access for human cohort studies:

RECOMMENDATIONS

1. Research funders should require explicit data sharing and management plans as part of grant applications (even if these plans conclude that data sharing is not appropriate); and should ensure that these issues are adjudicated before funding for new studies, or renewal of existing studies, is agreed. Funders should also review existing studies: where there is a prima facie reason to think data sharing might be desirable, the study should be contacted to clarify what policies are in place.

2. EAGDA funders should support work on: the development and dissemination of data formats and metadata standards; the promotion of interoperability across studies and datasets; and help to ensure that datasets are readily discoverable and accessible to potential users.

3. EAGDA funders should require study leaders to make data access processes transparent to potential secondary users of data, including: how datasets can be discovered and accessed; what requirements secondary users are expected to fulfil; and how decisions about access requests are made.

4. EAGDA funders should ensure studies establish proportionate governance procedures for data access, drawing on existing good practice and ensuring appropriate levels of independent input and oversight for access decisions, together with an independent appeals process.

5. EAGDA funders should be clear that using collaboration with the study team as the sole means through which to allow data to be accessed is not appropriate, other than in exceptional circumstances where it can be justified. In all cases, funders should expect the origins of data to be acknowledged in publications from secondary users.

Cont'd…

6. EAGDA funders should work with studies to ensure that the terms of consent enable data to be accessed and used in a way that maximises its value, whilst protecting the rights of research participants. It is critically important that the consent process explains clearly and as fully as practicable, how the data may be used in future.

7. EAGDA funders should set expectations that studies will develop clear policies on the management of depletable resources, ensuring guidance and support is provided to study leaders in this process.

8. Funders should work with research institutions and scientific journal publishers, to develop a proportionate system of adjudication and of penalties for breaches of data sharing rules, and guidelines for their use. EAGDA funders should share information on proven cases of data misuse.

9a. EAGDA funders should actively promote and evaluate initiatives to harmonise access processes and requirements for data users. Options include:
- Facilitating workshops of study leaders, administrators and data managers to explore options for harmonising data access processes;
- Developing equivalence of criteria for checking academic credentials when applying for data access for different studies, drawing on established examples of good practice and encouraging innovative work from the research community in this area;
- Supporting the development of uniform templates for access applications.

9b. EAGDA funders should jointly consider with institutions, study leaders and other key stakeholders (i.e. Research Ethics committees) whether consolidation across several study Data Access Committees would be an efficient, effective mod el for some studies, to ensure sustainability of access over time and independently of short-term competitive grant cycles tied to specific scientific projects.

10a. EAGDA funders should seek to establish the costs of data access activities for different studies:
- in the short term (setting up access mechanisms and covering data formatting and management costs); and
- in the long-term (ensuring sustainable storage, curation and access beyond the duration of a grant).

10b. EAGDA funders should also consider under what circumstances cost-recovery is an appropriate model for studies to operate for access to data and samples.

11. EAGDA funders should jointly discuss how data access and sharing infrastructure should best be supported, co-ordinated and sustained for the long term. They should agree clear expectations for study leaders on the use of established repositories with archiving facilities, in order to assure quality and ensure discoverability of data. Such repositories need to be sufficiently and sustainably funded, with clear definition of who is responsible for them.

EAGDA advises the funders to build upon the findings and recommendations of this report, working together with the research community across disciplines, to create a published UK framework of guiding principles and best practice for data access in research involving human participants.

ACKNOWLEDGEMENT

This is a report of the Expert Advisory Group on Data Access (EAGDA). EAGDA was established by the Medical Research Council, Economic and Social Research Council, Cancer Research UK and the Wellcome Trust in 2012 to provide strategic advice on emerging scientific, legal and ethical issues in relation to data access for human cohort and longitudinal studies across genomics, epidemiology and the social sciences.

The report is based on work undertaken by the EAGDA secretariat at the Group's request. The contributions of Natalie Banner, David Carr, Holly Baines, Katherine Littler and Joanna Scott at the Wellcome Trust are gratefully acknowledged. EAGDA would also like to thank the representatives of the MRC, ESRC and Cancer Research UK for their support and input throughout the project. Most importantly, EAGDA owes a considerable debt of gratitude to the individuals from the research community who contributed to this study through feeding in their expert views via surveys and interviews.

The Expert Advisory Group on Data Access

Martin Bobrow (Chair)
James Banks
Paul Burton
Rosalind Eeles
Mark Elliot
Paul Flicek
Mark Guyer
John Hobcraft
Tim Hubbard
Bartha Maria Knoppers
Mark McCarthy
Andrew Morris
Onora O'Neill
Melanie Wright

INTRODUCTION

Scientific progress has always depended on scientists releasing their data for others to inspect, for purposes of verification and to build a corpus of common knowledge. How much of the data are disclosed to others, and when this disclosure takes place, has varied greatly over time and between different scientific disciplines.

The past decade has seen an ever increasing move towards earlier and fuller sharing of research datasets. In the biomedical sciences, the Human Genome Project set a new paradigm for rapid pre-publication sharing. Subsequent international consensus statements have reinforced the view that "community resource projects" of this type, and other research datasets may have a wider utility than the specific goals and analyses envisaged by those who collected the data. The social sciences have a long history of data sharing, with well-established repositories, such as the UK Data Service and Inter-University Consortium for Political and Social Research, providing the infrastructure to preserve and access research datasets.

There are many good drivers of this policy move towards more open access to data, but it exposes important tensions:

- Where the research involves human subjects, they have extremely important moral and legal rights to protections, particularly of their privacy.
- The researchers who gather the data, having invested substantial time, effort and intellectual creativity in the process, also have important rights. The current incentives and rewards in the academic research sector do not adequately recognise the value of creating public data resources, and although the movement has been willingly embraced by many scientists, there is a need to align incentive and reward systems.[1]

Drivers for data sharing: The increasing expectation upon data producers to enable secondary access to the data resources they hold has been driven by a number of factors:

1. The scale of datasets being collected has grown dramatically, and these datasets are assembled at significant cost.
2. It will usually not be possible for one group to analyse these data exhaustively, and there will often be significant potential for the data to be used to answer questions distinct from the original research questions of the data producers.
3. Developments in information technologies are transforming the ease with which large datasets can be shared, linked and analysed.
4. Both those who volunteer their data and samples for research, and those who pay for that research, hope for progress towards useful and eventually applicable results for human health and other societal benefits to be as rapid as possible. Indeed, there is a clear ethical requirement for efficient use of data from human research participants.[2]

The ability to link cohort study datasets to information held in electronic patient health records and other population-based datasets also offers enormous opportunities for

[1] These issues were discussed in the EADGA report, 'Establishing incentives and changing cultures to support data access' (May 2014), see: http://www.wellcome.ac.uk/About-us/Policy/Spotlight-issues/Data-sharing/EAGDA/WTP056496.htm

[2] A recent report from the Institute of Medicine outlines guiding principles for responsible data sharing in clinical trials, championing data sharing as being in the public interest (January 2015): http://www.iom.edu/Reports/2015/Sharing-Clinical-Trial-Data.aspx

research that can lead to more effective, safer health services and treatments, and better understanding and new insights into diseases.

Concerns relating to rights of human subjects: These concerns arise within a wider context of increasing awareness about the way personal data of all types can be collected, tracked and used. In many cases the data involved are technically not "personal" data, in the sense that they are no longer individually identifiable. Nonetheless, the potential for data users to link datasets in order to triangulate information on specific individuals is increasing, and has in some rare situations enabled individuals to be re-identified from data that previously has been subject to anonymisation techniques. Risks of re-identification have generated particular concerns in genomics, and need to be carefully assessed.[3]

As sophisticated analytic technologies render the demarcation between personal and non-personally identifying data increasingly blurred, there is an ever greater need for careful governance over the handling of human data and the conditions under which they may be made available for access by researchers, even where the data being reused do not constitute personal data under the meaning of data protection law.[4] Consent is not a panacea and even where consent specifically allows for further data use, robust governance is essential for the ethical conduct of research.

There is a complex and shifting legislative and regulatory environment governing how personal data from human subjects should be protected. At the time of writing, there is on-going uncertainty over the provisions for research in the draft European Data Protection Regulation, and, despite the admirable intention of protecting the privacy of European citizens, there is potential that this Regulation could very seriously limit how datasets may be reused and shared for research.[5] At the same time, national bodies such as the Health and Social Care Information Centre are constantly reviewing their own policies regarding health record linkage. Whilst this ensures they can respond rapidly to short-term changes in public, political and societal perspectives regarding costs and risks associated with data access, it also leads to repeated shaking of the foundations of long-term research programmes which rely on those linkages being reliably available.

The need for common principles for data access: The data access process involves several stakeholders, all of whom have different rights, interests, risks and responsibilities with regards to the data (see **Annex 1**). Research groups in the UK have to date operated largely without strong policy guidance when setting up systems to negotiate data access. A number of different pragmatic governance solutions have emerged by which studies have sought to provide access to data users, while safeguarding participants and addressing the associated risks and concerns (see **Annex 2** for an overview of the UK funding and research environment).

EAGDA wanted to understand the data access landscape from different viewpoints, in order to provide policy advice in this area. This report focuses on the governance of data access for human cohort studies and seeks to set out a series of key principles and recommendations for funders, trying to strike this delicate balance. Although focused on the

[3] The risks associated with this issue have also been considered by EAGDA (October 2013): http://www.wellcome.ac.uk/stellent/groups/corporatesite/@policy_communications/documents/web_document/wtp055972.pdf

[4] The Nuffield Council on Bioethics recently considered the ethical issues associated with sharing health and biomedical data, and among their many insights of relevance to this report was the conclusion that legal compliance with regard to handling data is not sufficient to ensure data use is ethically appropriate (February 2015): http://nuffieldbioethics.org/project/biological-health-data/

[5] A statement of concerns about the draft EU Data Protection Regulation has been signed by a large number of non-commercial research organisations and academics: http://www.wellcome.ac.uk/stellent/groups/corporatesite/@policy_communications/documents/web_document/WTP055584.pdf

UK and directed at the EAGDA funders, we believe the principles represent best practice at an international level and we would welcome further discussion and co-ordination on applying these recommendations beyond the UK context.

EAGDA is very grateful to the EAGDA secretariat, who carried out some research on the existing situation. (see **Annex 3-7**). Based on the evidence collected, EAGDA concluded that:

- there is widespread support for the principle of early data access;
- there is some lack of clarity as to exactly what is expected of data producers – which studies should enable access to data, how much, how soon and under what restrictions;
- although the vast majority of larger studies have data access plans in place, the system as a whole is working inefficiently from the perspective of secondary researchers who wish to access and use data from different studies;
- funders have variable levels of oversight of the extent and manner in which their own grantees are implementing their policy requirements;
- because of the *ad hoc* nature of the development of governance mechanisms at this early stage of the shift towards enabling access to data, there is an opportunity for greater harmonisation and agreement on the definition of guiding principles and examples of best practice for data access and governance.
- A co-ordinated approach could foster efficiencies, and increase clarity for funders and researchers on what data assets are available and how to access them, without adding undue burden or cost.

We outline below the key areas in which we believe such good practice may be developed. We recognise that different studies raise very different issues and concerns, and we are not suggesting a one-size-fits-all approach. Rather, we are suggesting that funders should set clear expectations that studies should follow these common principles unless there are convincing reasons why it would not be appropriate to do so. These recommendations will be most relevant to large scale genomic, epidemiological, longitudinal and public health studies, although we consider that they represent best practice for all research involving data from human participants.

KEY PRINCIPLES

1. Determining which datasets should be made accessible

Many researchers are unclear as to which research datasets should be made accessible to secondary researchers. THIS judgement should take into account the size, complexity and generalizable utility of the data generated. The requirement to make data accessible is more generally applicable to population studies and large scale genomic, cohort or public health studies than to small scale laboratory experiments. There is a continuum of scale, which makes it difficult to set precise guidance. For studies that generate significant datasets, the presumption should be that a data access policy is put in place.

Where a recognised community data repository exists for a particular data type (such as the UK Data Service for social science data, or the European Genome-Phenome Archive (EGA) for genomic datasets), these should be used wherever feasible. Some repositories have specific data access requirements, which must be taken into account in planning.[6] The ESRC and the UK Data Service, which it funds centrally, is an exemplar of clear policy and

[6] For example, the EGA requires that all deposited datasets have a Data Access Committee of some sort. This could be the study PI or a fully constituted DAC making decisions; the EGA is not involved in access decisions but provides the storage and technical support to release data.

expectations being co-ordinated with sufficient funding and infrastructure, to motivate and support researchers to make data accessible.

Setting the threshold for access: Enabling data to be shared widely allows novel questions to be asked of the datasets that would not have been anticipated by the original study team. This can ensure efficient use of data resources by reducing costly primary data collection and making the most of datasets that have been collected, formatted and stored. There should therefore be a low threshold for allowing access to datasets by *bona fide* researchers, provided consent is adhered to and participant confidentiality is respected,[7] and there is adequate provision for recognition of and credit for those who were responsible for collecting the data.

Some data access requests require bespoke extractions and processing by the study team or data manager. Study leaders should be clear as to what resources are available for undertaking this work and set out a reasonable level of staff time and expertise for secondary users to expect to be available in supporting their requests.

In short, there should be an expectation that datasets will be made accessible unless there are good reasons to the contrary. Legitimate reasons could include concerns about inherent disclosure risk, for example, if data include high resolution geo-spatial data; if participants are from rare disease cohorts or a small geographical area; or if proposed linkages with other dataset pose a serious threat to anonymity or confidentiality.

The decision on accessibility needs to be made on a case-by-case basis. The access policy for each study should be made explicit and justified when submitting an application for funding, and this should be assessed by funders and their granting committees before funding is agreed.

Recommendation 1

Research funders should require explicit data sharing and management plans as part of grant applications (even if these plans conclude that data sharing is not appropriate); and should ensure that these issues are adjudicated before funding for new studies, or renewal of existing studies, is agreed. Funders should also review existing studies: where there is a prima facie reason to think data sharing might be desirable, the study should be contacted to clarify what policies are in place.

2. Making data discoverable and useable

It is important that potential users are able to discover the existence of data assets, and that high quality metadata is captured in a manner that facilitates data linkage and re-use. Many funders have made welcome progress in establishing accessible registers of data assets. The UK Data Service is notable in this regard for the support it offers researchers with high-value datasets in formatting data appropriately for storage, curation and archiving within its repository. The Digital Curation Centre based in Edinburgh also provides a wealth of guidance and support for researchers wishing to carefully curate their data, and initiatives such as this should be credited for their efforts in this technically complex area.[8] In addition to making data readily discoverable, longer term work on standardisation and harmonisation of formats for data and metadata will facilitate secondary data usage.

[7] For example: through anonymisation techniques; providing the minimum data necessary for the proposed use; and conducting statistical disclosure risk assessments where appropriate.
[8] Digital Curation Centre: www.dcc.ac.uk

Recommendation 2

EAGDA funders should support work on: the development and dissemination of data formats and metadata standards; the promotion of interoperability across studies and datasets; and help to ensure that datasets are readily discoverable and accessible to potential users.

3. Ensuring transparency

The process for seeking and gaining access to study datasets should be readily available to prospective users, together with details of: what criteria must be fulfilled when making an access request; what conditions there may be on access; how access decisions are made; who is responsible for those decisions; how to appeal against a decision; what costs might be incurred by the user; and what the approximate timescale would be for gaining access to the data.

Transparency has many advantages, both for the data producer and the user. For the producer, it ensures there can be independent scrutiny for the way access decisions are made, which may help protect the study from accusations of unfairly restricting access to datasets. Transparency also enables a degree of convergence in approaches to be sought between similar studies. For the potential data user, transparency enables easy discovery of how the application process works, which can help users quickly ascertain if making a request is likely to be an efficient use of their time and resources. For research participants, it helps to provide assurance in data governance processes and instil trust. To that end, providing public lay summaries of approved uses is of particular importance.

Recommendation 3

EAGDA funders should require study leaders to make data access processes transparent to potential secondary users of data, including: how datasets can be discovered and accessed; what requirements secondary users are expected to fulfil; and how decisions about access requests are made.

4. Handling of data access requests

Some studies handle data access requests through the study leader or the study team. This may be entirely appropriate in some cases, for example in smaller studies where the number of access requests is anticipated to be low. For larger studies, it is common to institute some form of Data Access Committee (DAC) to adjudicate individual requests for access to the data.

A DAC is an advisory group with expertise independent of the study leader and research team, which is set up by the study leaders to: safeguard research participants and data producers; help with the management of data; and provide transparency to decision-making and access decisions. DACs do not have legal standing as data owners or curators, but have delegated authority to make decisions on behalf of the study team.

The balance of expertise required on a DAC may comprise individuals with scientific, technical, legal, statistical and ethics expertise. It should be chaired by a member independent of the study team. It may be desirable for a study team member familiar with the datasets to review the appropriateness of access requests, although such individuals should act in an advisory role.

DACs should ensure that access systems function efficiently and with proper regard to the interests of study participants, data producers, funders and secondary data users, using their expertise to determine and balance the risks involved in different potential levels of disclosure. This can be both on a case by case basis and on broader matters of data access policy for the study.

Virtually all DACs take some or all of the following issues into consideration for each application:

- Ensuring that secondary users are *bona fide* researchers. Those who donate their data for research probably expect that the research will be carried out by those with demonstrated competence in the field. The research is also more likely to be useful and productive if carried out by someone with adequate research experience, and there are accepted (although not always enforceable) standards of good conduct amongst professional researchers.
- Ensuring the terms of participant consent are adhered to: (see section 6 below).
- Protecting the confidentiality and other interests of participants. DACs may undertake assessments of disclosure risks and it is common to prohibit secondary data users from attempting to re-identify or re-contact research subjects.
- Ensuring that the rights of those who worked to produce the data are appropriately acknowledged and respected: it is usual to ask for undertakings on the nature of timing of research publications, and acknowledgements of data sources, designed to protect the rights of the primary researchers responsible for generating the data.
- Ensuring that there is evidence of a responsible organisation standing guarantor to the appropriate use of the data. Having applications countersigned by a responsible employing authority ensures that some substantive body or organisation will take responsibility if rules of access are found to have been broken.
- Ensuring compliance of secondary use with any legal and regulatory requirements, such as data protection legislation.
- Attempting to avoid inadvertent duplication in uses of datasets by different groups.
- Evaluating the scientific merit of access requests. It is not generally the business of a DAC to assess the scientific merits of applications for secondary data usage. In some cases however, it may be clear that a proposed investigation cannot be successful, and it is appropriate to draw this to the applicants' attention.
    - There may be grounds for refusing applications thought likely to bring the main study into disrepute, for example, if the applicants are attempting to investigate a contentious topic in a way which cannot for scientific reasons be supported by the available data. In general, however, since the data are not a depletable resource, there should be a very low threshold to allowing access. This is entirely reversed in the case of requests for use of depletable resources such as samples are requested (see section 7 below).

Applicants for data use are normally asked to complete a short application covering the preceding points, and an agreement (countersigned by their employing authority) setting out their agreement to the terms of data usage.

DACs do not necessarily need to consider all requests to access a data resource. For example, routine requests for low-risk data may simply require proof that the researcher is *bona fide* and may be approved through administrative processes without referral to the DAC. In other situations, third-party data repositories (such as the UK Data Service in the social sciences) may process access requests on behalf of the study for the distribution of a

set of pre-agreed data files under pre-agreed procedures. In both situations, study DACs only need deal with requests for particularly sensitive data (typically biomedical data or highly disclosive variables).

An important component of having a transparent and accountable access process is that those who request data be able to challenge a decision to refuse or restrict access. Such appeals should be overseen by an independent adjudicator or board not involved in the original access decision. These appeals channels may rarely be required in practice, particularly if the process of data access is clear, prospective users are able to understand what criteria they need to fulfil to make a request, and there is a channel of communication between the study managers and potential data users.

Recommendation 4

EAGDA funders should ensure studies establish proportionate governance procedures for data access, drawing on existing good practice and ensuring appropriate levels of independent input and oversight for access decisions, together with an independent appeals process.

5. Establishing fair conditions for access.

Some studies only permit secondary use of datasets on the condition that a member of the study team collaborates on any analysis and publication. There may on occasion be valid reasons for this approach, for example, some complex datasets require a thorough understanding of the methods of collection in order to avoid misrepresentation or misinterpretation in analysis. However, such restrictions may prevent novel or innovative uses of the data, and may be perceived as a means of limiting access. Whilst data access in the context of collaborations should be encouraged, collaboration should not be the sole means through which a study allows access to its data unless there are explicit and very strong reasons for this approach, accepted by funders and their adjudicating committees.

Conditions are sometimes imposed on users of datasets, such as the expectation that the study leader or study members are listed as co-authors on subsequent publications, irrespective of whether or not they have made a contribution to the paper beyond supplying the data. It is legitimate for studies to track how their datasets are used, and to expect appropriate acknowledgment of the origins of the data, but co-authorship should recognise significant contributions to a publication and not be a default requirement for permitting access to data.

Although it is common practice for data producers to assist and advise secondary users on how to make the best use of their datasets, the quality of secondary analysis undertaken on datasets is not the responsibility of the study team and should only be grounds for refusing access if there is a real risk of damage to the study, for example, by potentially losing the trust and on-going support of participants in a longitudinal study. In any of these cases, demonstrable concerns need to be transparently articulated by those making decisions about access.

Recommendation 5

EAGDA funders should be clear that using collaboration with the study team as the sole means through which to allow data to be accessed is not appropriate, other than in exceptional circumstances where it can be justified. In all cases, funders should expect the origins of data to be acknowledged in publications from secondary users.

6. Respecting the terms of consent

The great majority of research datasets involving data collected from human subjects are based on participants explicitly consenting to take part in research. This consent represents an agreement between those donating their personal data for research and the research community, and should always be meticulously honoured in order to maintain the trust of the participants in the research process and its governance.

The nature of this process of obtaining consent has been examined in depth and there are a number of different models of consent that may be appropriate in different circumstances. [9]

Current best practice is for participants to be given explicit information about the likelihood of both summary and individual-level data being made available for secondary research, the possibilities for data linkage, possibilities for scientific collaborations across national boundaries, cloud storage of data or other relevant issues as appropriate to the individual study, and how access will be governed. Specific provisions must be in place to safeguard subject privacy, and the limitations of these safeguards need to be made clear during the consent process.

Many different options may be available, allowing subjects to delegate greater or lesser degrees of control over the data to the study leaders or other governance groups. Particular secondary uses (for example, research on race or intelligence) or users (for example, commercial organisations) should only be excluded from terms of consent if explicit justifications for such restrictions are agreed at the funding application stage. Good practice would attempt to ensure that the data, once collected, has the greatest possible utility consonant with the wishes of the participants and the provisions of data protection legislation.

In dealing with legacy collections, however, the terms of consents given historically must be respected, and DACs have a duty to ensure these are not breached. Where there is ambiguity in interpreting precisely what forms of research or access fall within the consent, an appropriate research ethics committee or other similar body should be consulted rather than the primary researchers taking responsibility themselves for this decision. In some situations consideration can be given to re-contacting and re-consenting the study participants. The use of de-identified data for purposes not clearly within the original consent can sometimes be justified, where re-consenting is not a practicable option, but these decisions must be taken by an independent ethics committee and not by the researchers nor by the DAC advising them.

Recommendation 6

EAGDA funders should work with studies to ensure that the terms of consent enable data to be accessed and used in a way that maximises its value, whilst protecting the rights of research participants. It is critically important that the consent process explains clearly and as fully as practicable, how the data may be used in future.

7. Managing depletable resources

Unlike data, biological samples from study subjects are depletable resources. A key consideration for any access policy is how to make best use of them. Good stewardship requires that the resources are used for the highest quality science, as best this can be

[9] Human rights approaches, such as those adopted by the Global Alliance for Genomics and Health, may provide a supportive framework for widening the scope of governance beyond a narrow focus on informed consent.

assessed. The judgement is particularly difficult because current applications must be judged against unknown potential future applications.

Ascertaining how depletable resources should be managed will be highly dependent on the particular study. Study leaders may require independent expertise to assist the study team, given that it is often not possible to anticipate how advances in technologies and science could alter what the best use of the samples could be in the future. Studies should develop a policy for how depletable resources will be strategically managed, in order to provide clarity to potential secondary users (and participants). For example, larger studies could issue specific calls for proposals, enabling access requests to be weighed against one another to decide how best to use the resources.

Recommendation 7

EAGDA funders should set expectations that studies will develop clear policies on the management of depletable resources, ensuring guidance and support is provided to study leaders in this process.

8. Sanctions for misuse

Serious breaches of Data Access and Material Transfer Agreements, such as deliberate attempts to re-identify research participants, should be treated seriously by studies, funders, publishers and institutions. Using appropriate sanctions in cases where allegations are proven can encourage careful and thorough data protection practices and can act as a deterrent against breaches.

Sanctions could include:

- withdrawal of access and prevention of future access to any study datasets for the non-compliant researcher or in case of a serious breach, their entire institution.
- funders withdrawing access to current or future funding.
- forcing the withdrawal of any manuscript or published paper based on misused data.

Sanctions should be proportionate to the scale of the offence and target the appropriate level of the breach; an institutional failure to follow good data security practices should be treated differently to an individual researcher seeking to re-identify a research participant for some personal gain.

If sanctions are to be imposed, an independent adjudication process for serious penalties is necessary. Our understanding is that sanctions are very rarely needed, but a process should be in place for handling these issues transparently and fairly when they arise, with penalties spelled out clearly in advance to researchers. In order to be able to enforce sanctions when needed, compliance with data access agreements needs to be monitored in a proportionate manner.

Recommendation 8

EAGDA funders should work with research institutions and scientific journal publishers, to develop a proportionate system of adjudication and of penalties for breaches of data sharing rules, and guidelines for their use. Funders should share information on proven cases of data misuse.

9. Establishing effective and harmonised data access governance systems

A major theme reported by those seeking access to data is the need to pass the same access hurdles in multiple, slightly different forms, as access to each dataset must be

separately negotiated. This is reminiscent of the situation that used to pertain in the UK and elsewhere with regard to Research Ethics Committees (RECs), when each local institution had its own REC with separate rules, forms and norms. Much effort is wasted, for little obvious gain in either ethical stringency or protection of research participants.

Although local characteristics and the context of each study need to be taken into account, many research studies have strong commonalities in their requirements for data access applications. Consistent or co-ordinated approaches between different studies may be more efficient for the studies themselves, as well as simplifying access processes for potential applicants, and should be sought where possible. These may utilise technological solutions such as globally agreed identifiers for individual researchers, for example based on the ORCID system.[10]

Different jurisdictions may impose different regulatory requirements on suppliers and users of data, creating obstacles to data access operating across borders. For fields such as genomics that may require a large number of datasets and global levels of co-operation, the challenge of harmonizing access policies and procedures internationally is substantial. Organisations such as the Global Alliance for Genomics and Health are working towards understanding the legal, practical, technical and ethical issues of global data sharing in genomics and EAGDA considers this a valuable initiative.[11]

A range of options are open to studies, from simply using the same application forms and processes as other similar studies, to sharing DACs and having common access processes enabling a single application to handle access to several different datasets. The ACCC (Access Committee for CLS Cohorts) provides one historical example whereby access requests for three different cohorts within the Centre for Longitudinal Studies that require discussion were handled by the ACCC, whilst the individual studies themselves handled routine requests for data and managed the logistics and administration of the access process.[12] Other examples include UK10K which historically used the same access process as the Wellcome Trust Case Control Consortium (WTCCC);[13] and DbGaP which handles access requests for all NIH-funded research datasets on genotypes and phenotypes.

Before setting up a study, study leaders, institutions and funders should work together to consider whether an existing DAC could be utilised or adapted to consider access requests for the planned study. When considering the appropriateness of consolidation, the funding base for such DACs will need to be clarified among the constituent studies and their funders.

Consolidation may provide advantages in terms of sharing resources, expertise and best practice, but may also generate challenges in terms of establishing the high degree of collaboration and harmonisation that may be necessary to bring access decisions under one DAC, and requiring a substantial degree of initial co-ordination among studies, institutions and the relevant funders to set up. Consolidation will only be appropriate where it generates efficiency savings and simplifies decision-making: it is intended to facilitate smooth data access and not to overcomplicate the process either for data producers or for users.

Recommendation 9a

EAGDA funders should actively promote and evaluate initiatives to harmonise access processes and requirements for data users. Options include:

[10] ORCID provides a persistent digital identifier for researchers: http://orcid.org
[11] Global Alliance for Genomics and Health: http://genomicsandhealth.org/ . Three members of EAGDA are involved in this initiative.
[12] St the time of writing, this mechanism is being reviewed with a view to further consolidating the access mechanisms with other studies.
[13] At the time of writing, access requests for UK10K data are handled by the Wellcome Trust Sanger Institute's Data Access Officer.

- **Facilitating workshops of study leaders, administrators and data managers to explore options for harmonising data access processes;**
- **Developing equivalence of criteria for checking academic credentials when applying for data access for different studies, drawing on established examples of good practice, and encouraging innovative work from the research community in this area;**
- **Supporting the development of uniform templates for access applications**

Recommendation 9b

EAGDA funders should jointly consider with institutions, study leaders and other key stakeholders (i.e., Research Ethics Committees) whether consolidation across several study DACs would be an efficient, effective model for some studies, to ensure sustainability of access over time and independently of short-term competitive grant cycles tied to specific scientific projects.

10. Meeting the costs of data access

Funders, and the research community more broadly, lack granular information about how much data access activities cost. Costs are currently either absorbed into core grant costs or use a cost-recovery model, either charging a flat fee for all requests or tailoring fees to the specific data request based on a calculation of the staff time and resources required to extract the necessary data to fulfil the request.

Costs relating to data access include the formatting, documentation and storage of data, staff expertise for processing requests and dealing with queries, administration of data access requests and the costs of running an access committee where needed. For some datasets, and some access requests, the amount of specialist time required to supply the data in a useful format can be considerable.

Researchers are often asked to submit data management plans as part of grant applications, but these are not always separately costed and it is rare for specific funds to be allocated to data access processes during the funding process. While we would not wish to see a new bureaucracy grow around this issue, if funders wish to encourage good governance and infrastructure for data access, some work will need to be done to estimate real costs, so that guidance can be developed for researchers. Funders should be prepared to meet reasonable costs for formatting, access mechanisms and curation, and to undertake some light touch audit of data access for studies once up and running, if they wish to maximise the value of the datasets generated.

Where data are stored in repositories, such as the UK Data Service or EGA, significant administrative, storage and curation costs are borne by the repositories themselves and their funders. Overall, however, the use of repositories will be very cost-effective in providing for long-term sustainable and harmonized access. Indeed, sustainability is a key issue, since many research studies have a fairly short time span of operation and it is important that arrangements are in place for the curation and control of the dataset when the research team have dispersed.

Discussions around costs are often framed as a choice between core-funding of a study or studies operating a cost-recovery model, but it needs to be recognised that this is not a “competition” between the data provider and the data user for covering costs. Rather, it is a question of whether the funding for data access should be funded entirely by the funding agency that creates and supports the data producer, entirely by the funding agency supporting the secondary research to be undertaken using the data, or somewhere in

between. This requires a debate among repositories, institutions, researchers and funders, with acknowledgement of shared responsibility between the funders – and it would seem likely that the ultimate position should lie "somewhere in between".

Recommendation 10a

EAGDA funders should seek to establish the costs of data access activities for different studies:

- **in the short term (setting up access mechanisms and covering data formatting and management costs); and**
- **in the long-term (ensuring sustainable storage, curation and access beyond the duration of a grant).**

Recommendation 10b

EAGDA funders should also consider under what circumstances cost-recovery is an appropriate model for studies to operate for access to data and samples.

11. Sustainability of repositories

Several generic data repositories are well-established as key facilities for depositing data from different research fields (such as the UK Data Service and EGA, and dbGaP in the USA). In the UK, the UK Data Service has a mandate and is funded to store, manage and make available data from relevant studies funded by its supporting agency (the ESRC). The EGA also has a strong reputation across a broad international field, but receives data from many different sources without a specific central mandate. If other funders wish to ensure their researchers make data available for secondary use through shared, centralised repositories, they would need to make this clear to researchers from an early stage of the grant awarding process. The funding of such repositories need to be sustainably managed for the long-term.

Discussions between researchers, funders and institutions are needed to establish where responsibility for sustainable data access lies and how activities can be co-ordinated on a large scale to support data repositories sustainably. Funders and institutions should co-operate to ensure that their data sharing policies do not conflict: researchers might be required by their university to store data in an institutional repository, which may be acceptable to funders provided the datasets are discoverable via a platform (such as the MRC Research Data Gateway for population health datasets) or other well-established portal and accessible under clear governance arrangements. Whatever the arrangements, these need to be clarified between the different stakeholders and kept as simple as possible for the research community.

Recommendation 11

EAGDA funders should jointly discuss how data access and sharing infrastructure should best be supported, co-ordinated and sustained for the long term. They should agree clear expectations for study leaders on the use of established repositories with archiving facilities, in order to assure quality and ensure discoverability of data. Such repositories need to be sufficiently and sustainably funded, with clear definition of who is responsible for them.

Cancer Research UK
The Angel Building
407 St John Street
London EC1V 4AD
T +44 (0)20 3469 8360
E publicaffairs@cancer.org.uk
www.cancerresearchuk.org

Economic and Social Research Council
Polaris House
North Star Avenue
Swindon SN2 1UJ
T 01793 413000
E comms@esrc.ac.uk
www.esrc.ac.uk

Medical Research Council
Polaris House
North Star Avenue
Swindon SN2 1FL
T 01793 416200
E corporate@headoffice.mrc.ac.uk
www.mrc.ac.uk

Wellcome Trust
Gibbs Building
215 Euston Road
London NW1 2BE, UK
T +44 (0)20 7611 8888
F +44 (0)20 7611 8545
E contact@wellcome.ac.uk
wellcome.ac.uk