Statement for EAGDA funders on re-identification

The Expert Advisory Group on Data Access (EAGDA) is a group of experts established by the ESRC, MRC, CRUK and the Wellcome Trust to provide strategic advice on emerging scientific, ethical and legal issues in relation to data access for studies across genetics, epidemiology and the social sciences.[1]

THE ISSUE

1. Every reasonable effort must be made to protect the privacy of research participants. In consultation with experts in the field, EAGDA are currently examining the issue of potential re-identification of anonymised individual research subjects from genomic and other data in the UK (including research, administrative and public record data). We have also sought advice from the Information Commissioner's Office.

2. Some participants in the "1000 Genomes" Project have recently been identified by combining publicly available demographic information with their anonymised genomic data.[2]

3. It is now clear that although the data in a genomic dataset may be fully anonymised in the conventional sense, cross-linking with general demographic data that are available from elsewhere makes it technically possible in some circumstances to triangulate the identities of individual research participants. Large datasets, particularly those including extensive genomic information, cannot be completely safe from inferential exploitation, including subject re-identification. Although the likelihood of such re-identification may currently be low for most types of study, it is likely to increase in the future as:
   - Research datasets become richer, more complex and more readily accessible;
   - Methods of analysis and interpretation increase in sophistication and reduce in time and cost;
   - Improved and wider use of demographic and administrative data become possible; and
   - Individuals release more information about themselves into the public domain e.g., through recreational genomic and social networking web sites.

4. Proper consideration of all the relevant issues in this is very rapidly evolving field will take some time. We therefore consider it timely to provide interim advice to funders and researchers on the risks of re-identification, its potential impact on research practices and possible mitigating steps. We wish to promote discussion among relevant research communities about ways to manage these risks without creating unnecessary impediments to sharing data for research. The types of issues that need to be considered are:
   - Whether to control access to particular data items or classes of data;
   - How to maintain such data securely;
   - How to ensure good data governance – e.g., who should be able to access what data, from what type of repository and via what access mechanisms, subject to what conditions and obligations of use; and
   - What penalties might be applicable to those who breach data access agreements.

RECOMMENDATIONS

5. Good data governance ensures the utility of datasets can be maximised while ensuring participants' privacy is respected. We recommend that the following good practices are implemented by researchers and funders:

Consent

6. Study participant consent processes should include explicit discussion of the risk of re-identification and should acknowledge that there may be future risks that are currently unanticipated. Although each project will vary in detail, examples which we consider to be good practice can be seen in the HapMap[3] and 1000 Genomes project[4] consent templates, which set out such risks but rightly emphasise that they are currently small.[5]

Risk assessment

7. An assessment of the risks of re-identification should be undertaken as part of the planning of any new study, and reviewed at intervals thereafter. Established studies should take such steps now. Full consideration of the wider data environment is necessary when undertaking risk assessments to determine access controls on a dataset. A proportionate approach to data access and sharing needs to properly recognise the risk of re-identification, but must weigh this against the societal benefits of allowing wide access to data for secondary research.

Access control

8. It is already standard practice that data known or believed to be potentially identifying are kept on controlled access databases such that access is subject to formal data access agreements, and we fully endorse this. Nonetheless, researchers should recognise that classes of data that we currently believe to be safe might later be shown to inferentially disclosive. Policies should therefore be in place to enable rapid, flexible action to combat new threats to re-identification, without unduly damaging the scientific utility of research data.[6]

Data access agreements and sanctions

9. Malicious breaches of research data security are extremely rare, and there is no reason for researchers using data for agreed purposes to be alarmed. Nevertheless, it should be made clear as a matter of best practice that there are legal as well as scientific penalties for breaching the privacy of participants. This will provide reassurance to researchers and participants that data security is taken seriously and there are robust systems in place to deter against malicious use of data. Data access agreements should clearly set out both legal (see below) and funder-imposed sanctions for attempting re-identification from research datasets.

10. Funders' sanctions could include withdrawal of funding or access to data resources for individuals and their institutions if re-identification has been attempted or allowed to take place. It is funders' responsibility to provide clear notice of the sanctions process, including a statement of who will make these judgements and on what evidence, together with a transparent appeals process.

11. We have worked with the ICO to produce the following statement, designed to provide clarity to researchers as to the existing legal sanctions against re-identification:

"The Data Protection Act (1998) provides that personal data must be processed fairly and lawfully. Deliberately attempting to re-identify individuals from anonymised research data is likely to be unfair to the data subjects. This does not prohibit the use of personally identifiable or coded data for research, as the DPA contains a limited but important exemption (s.33) for research purposes. The exemption allows for personal data to be used, provided that:

a) identification does not take place; or,
b) if identification does take place, research participants have not previously been assured that only anonymised data would be published from the study (ICO Anonymisation Code, p.21).

Deliberate re-identification of individuals from anonymised research data, without express consent of the research participant, is likely to lead to a breach of data protection principles and could in turn lead to penalties enforced by the ICO, including a civil monetary penalty of up to £500,000. If data are transported or transmitted outside of the UK and subsequent re-identification takes place, the data controllers responsible for the data within the UK may be held liable for the breach if they have not taken adequate steps to protect against it. If this re-identified data is then subsequently obtained, recorded or held in the UK, the holder also would be subject to the principles of the DPA and potentially the penalties described. The ICO will treat deliberate re-identification activity with the utmost seriousness."

This statement helpfully clarifies not only that the act of re-identifying anonymised subjects without their consent is in itself potentially subject to penalties, but also that the use of such data to cause harm to data subjects (e.g. by discriminating against them) may also be subject to penalties even if the actual re-identification occurred outside UK jurisdiction. This may serve to reassure both research subjects and researchers that, although re-identification is becoming technically feasible, it is not an acceptable practice within the UK and may be subject to legal action.

12. There is a need for public dialogue to address the risks of re-identification and other possible consequences associated with the use of genomic data. Funders should pursue opportunities in this area. As the Government's 100,000 Genomes project is in development, it is timely to consider ways to undertake public engagement on disclosure risks and/or the benefits of data sharing for research and clinical practice.

13. EAGDA is continuing to work on this subject and will provide further advice to studies and funders in due course. We welcome comments and suggestions from interested parties.

---

[1] http://www.wellcome.ac.uk/About-us/Policy/Spotlight-issues/Data-sharing/EAGDA/index.htm
[2] M. Gymrek et al (2013) *Science*, 339; 321.
[3] p. 3 http://tinyurl.com/q6pzh8w
[4] Sections 7-8: http://tinyurl.com/nz652gu
[5] Both templates indicate potential methods of re-identification, but these may become rapidly outdated. We recommend specific details of possible methods is not given in the information to participants, as new techniques and methods will inevitably develop in future.
[6] The lack of such policies led to a significant, though temporary, disruption of the release and use of genetic data after publication of an influential article revealing it was possible (under some circumstances) to determine whether a particular individual was or was not represented in a particular set of aggregated genomic data, Homer, N. et al (2008) *PLoS Genet* **4**(8) doi:10.1371/journal.pgen.1000167