MINUTES OF THE FOURTH MEETING OF THE EXPERT ADVISORY GROUP ON DATA ACCESS (EAGDA)

10.30-16:00 MONDAY 14 OCTOBER 2013, WELLCOME TRUST

Present:

James Banks
Martin Bobrow (Chair)
Paul Burton
Rosalind Eeles
Paul Flicek
Tim Hubbard
Bartha Knoppers
Mark McCarthy
Andrew Morris (by phone, items 1,2,3)
Nigel Shadbolt
Chris Skinner
Melanie Wright

Geraldine Clement-Stoneham – MRC
Vanessa Cuthill- ESRC
Helen Dewberry – ESRC
Fiona Reddington – CRUK
David Carr- Wellcome Trust
Katherine Littler – Wellcome Trust
Natalie Banner – Wellcome Trust

Also present for morning:
Mark Bale – Genomics England
Laura Riley – Genomics England
Mike Parker – Ethics Advisory Group for Genomics England
Peter Mills – Nuffield Council on Bioethics
Martin Richards – Nuffield Council on Bioethics
Nicola Perrin – Wellcome Trust (item 1)

Apologies for absence:

George Davey Smith
Onora O'Neill
Mark Guyer
Peter Knight (new member)
John Hobcraft – ESRC
Peter Dukes – MRC

1. Updates

Minutes from the March EAGDA meeting were ratified.

The addition of Peter Knight from the Department of Health to the EAGDA was noted and welcomed.

Updates from funders were postponed to the afternoon session.

Nicola Perrin provided an overview of developments in clinical trials regulations and took questions. Key points addressed were:

- Clinical trials transparency is an issue that overlaps with EAGDA's interests.
- Momentum towards creating transparency has increased in light of Ben Goldacre's book 'Bad Pharma' and the AllTrials campaign.
- The practicalities of achieving transparency are difficult to establish, particularly in terms of meeting obligations to trial participants and ensuring confidentiality is maintained while enabling access to patient-level data.
- There have been recent movements towards the active policing of compliance: the HRA will now require trials to be registered in order to gain Research Ethics Committee approval.
- Several issues remain to be worked out, around the commercial sensitivity of trial data, and the possibilities for reinterpretation of data by different parties.
- GSK and Roche are developing a controlled-access model for registering trials and depositing trial data that could be expanded to a global initiative.

ACTION: The Secretariat to circulate a copy of the joint funder response to the EMA consultation on clinical trials from May 2013.

ACTION: The Secretariat to keep a watching brief on developments in clinical trials regulation and proposed access mechanisms to trial data.

2. The Open Data Institute (ODI)

Nigel Shadbolt provided an overview of the aim, role and scope of the ODI, which was launched one year ago with government investment, to create a demand site for open data. Key points of interest were:

- The ODI has global political support and international interest, and aims to create both economic and social value through the use and repurposing of a wide range of data sources.
- The G8 "open data charter" supports the creation of a range of national data as open data, and aims to engage public, private and third sector organisations to create a stable economic model of data sharing.
- Climate change data is very advanced in terms of creating interoperability of standards, machine readable data and the use of high resolution datasets for different purposes. This field may provide useful lessons for developing interoperability for health and medical data.

- The ODI is working to develop both supply and demand sides of the market for data: producers need to consider how to create demand and ensure data they invest in producing, formatting and making accessible will be valuable to others.
- Improving transparency and accountability are central to the ODI's aims. If companies (or funders, institutions or studies) have transparent and accountable processes this strengthens the argument for light-touch regulation.
- The ODI has developed a self-certification standard as a set of principles ensuring best practice for data producers to assess the quality and integrity of their data (covering aspects such as metadata, sampling methodology, granularity of data and accessibility). This may be further developed into an accreditation standard over time.
- An early demonstration of potential has been shown by a start-up company examining prescribing patterns of statins across Primary Care Trusts in England. They identified a potential saving to the NHS of £200m in six weeks if prescriptions for proprietary statins were replaced with generics.

3. Nuffield Council on Bioethics (NCOB) – Biological and Health Data inquiry and report

Peter Mills and Martin Richards introduced the working party and planned public consultation (due out this week) on exploring the harms and benefits that may arise from using (or not using) biomedical data, and the biomedical uses of other types of data (for example, administrative or social science data).

The group is working to identify the relevant values, norms and underlying practical concerns around the use of data and is planning to commission research into the actual harms that may result from security breaches, infringements of privacy, and their prevalence. Aiming to provide conceptual clarity and develop an ethical framework, the report is due to be published late in 2014. Discussion included:

- The potential implications of new methods for using and linking data for existing models of research and data access governance.
- That any consideration of potential harms ought to include the harms of not sharing and using the data.
- The need to consider data use and sharing as cross-jurisdictional, and what the ethical, social and legal implications of this might be.

There was consensus that the potential for joint work between EAGDA and NCOB on aspects of commissioned research for the report should be explored further. An analysis of the actual risks and harms of using health data, the harms from not using it, potential benefits and the prevalence of security breaches could inform EAGDA's work on identifiability and quantifying risks, particularly with regard to setting boundaries for controlling and accessing data.

ACTION: The Secretariat to circulate NCOB's slides and to follow up with NCOB to discuss options for working together.

4. 100,000 Genomes Project

Male Bale introduced the 100,000 Genomes project and the company that has been set up to run it, Genomics England. The project is being pursued now as the cost of sequencing has decreased significantly in recent years, and it is important to begin to understand

diseases as usually involving more than one gene. This cannot be done incrementally in the NHS.

- The project will strengthen the UK's reputation and expertise in genomics and sequencing technology, driving the improvement of data, linkage and annotation tools. It also aims to drive up skills in the NHS, particularly in informatics.
- The CMO Dame Sally Davies will take oversight of the science and public trust in the initiative.
- Genomics England is a company limited by guarantee owned by the Department of Health.
- It is aiming to move forward with pilots in 2014, progressing to clinical grade whole genome sequencing programmes by May 2015, and enabling sequence data to be linked to patient records.
- For the pilot programmes, 2000 genomes from patients with rare diseases, and 3000 genomes from cancer patients and their tumours, will be sequenced.
- There will be competition for the procurement of sequencing technologies. UK contracts will be preferred to ensure trust and that data are not distributed around the world.
- Public Health England is looking into the use of whole genome sequencing for tuberculosis and hepatitis C; Genomics England will be looking to develop partnerships with such organisations.
- Preliminary work on social attitudes and perceptions indicated that although many people were unfamiliar with the concept of 'genomics', there was strong support.
- Although the project is ambitious, there will be a first mover advantage to having genomic data in health systems and built into infrastructure; the research benefit is secondary.

Laura Riley (Ethics Advisor for Genomics England) and Mike Parker (Chair of Ethics Advisory Group for Genomics England) outlined some of the ethical issues arising from the project and the work that has been initially undertaken:

- Ethics advice and support will be built into the workings of Genomics England, including issues around consent and feedback to patients, benefits to the NHS, and data sharing. The Ethics Advisory Group includes patient representatives.
- The pilot phase of the project is an opportunity to learn about the ethical issues that will arise in practice, to allow a model of good practice and patient information leaflets to be developed, and an evidence base to be generated, for the main phase of the project.
- Ethicists and social scientists will be involved in research with participants in the pilot phase, focusing on consent in the first instance to enable the project to proceed.

The group discussed the initiative and raised the following issues:

- It is unclear whether there will be clinical benefit in the short term from whole genome sequencing.
- It will be important for the quality of the science to carefully determine the disease cohorts to be sequenced, although this will need to be balanced with the pragmatic issue of which NHS Trusts sign up to the project.
- A climate of data conservatism (in light of the recent international data privacy breaches) may prevent data from being shared, which will be to the long term detriment of the utility of the data: value could be compromised by overcautious safeguarding concerns.
- It is not yet clear who will have 'ownership' of the sequencing data.

- The power of whole genome sequencing will be in its linkage with clinical data, but clinical data are frequently compromised by poor quality and incompleteness. Efforts to produce clinical grade sequences will need to be matched by steps to improve and standardise clinical data.
- Clarity and openness is needed about the primary purpose of the project and how research objectives will fit into its aims. The availability of data, nationally and internationally, and the possibility of commercial use will also need to be clarified in order to maintain public trust.

ACTION: EAGDA to maintain a watching brief on developments with the 100,000 Genomes project.

5. Global Alliance

Martin Bobrow provided an introduction to the creation of the Global Alliance, with further comments on ethics and law by Bartha Knoppers, and data security by Paul Flicek. These three EAGDA members are members of the interim Global Alliance steering committee. The Alliance is in the process of being set up, with the intention of creating both public knowledge and commercial benefit from the combination of genomic analysis and clinical data.

There is substantial genetic variance with the most commons diseases and extremely large sample sizes are needed to determine the small contributions from large numbers of loci on disease variance. This is only possible if healthcare organisations and research studies from across the world are able to organise, pool and interrogate each other's' data.

The Alliance aims to work towards interoperable data and ethical standards and standards of behaviour to enable this scale of sharing and meaningfully work with others globally.

- Several UK organisations including CRUK, the Wellcome Trust, Department of Health and some universities have signed letters of intent in support of the Alliance. The Alliance now has in excess of 150 signatories.
- The movement is now towards developing Memoranda of Understanding, which would commit organisations to a firmer idea of the Global Alliance's intentions.
- Ethically and legally, the Alliance hopes to create a basis for collaborative working across different jurisdictions, through international codes of conduct or flexible standards on issues such as consent, data access and privacy.
- Technically, the Alliance hopes to allow searches to be conducted and datasets to be interrogated irrespective of the physical location of the data servers.
- Harmonisation of detailed phenotype information will be essential to this initiative, in order to render the genotype data meaningful for researchers and clinicians.
- The linking of environmental variables to genotype an phenotype data (requiring more than just a large sample size) was acknowledged as an important factor to consider, but the focus on linking genotypes to phenotypes is considered to be a first step for some important questions about disease.

ACTION: The EAGDA members who are closely involved with the Global Alliance to keep EAGDA informed of any relevant developments

6. Funder Updates

Vanessa Cuthill updated the group on developments with the ESRC's activities in relation to data access and sharing;

- The ESRC has been active in three areas:
  - Enhancing cohort studies
  - Working with business data to explore linking social science and commercial datasets
  - Setting up the Administrative Data Research Network, opening up de-identified routine administrative data for academia and the government (potential for access by commercial users will be considered at a later stage)).
- ESRC are in discussion with the ONS over the future of the census and administrative data collection, in light of budget cuts.

Geraldine Clement-Stoneham updated the group on MRC's recent activities:

- The four e-Health Information Research Centres have been launched
- A review of cohort studies will be discussed at an upcoming strategy meeting.
- The MRC are working on safe data and safe havens with international collaborators including Science Europe.
- The MRC data sharing gateway and service are being developed to provide a directory of information about cohorts.

Fiona Reddington updated the group on CRUK's recent activities:

- CRUK is looking into rates of compliance with its data sharing policies.
- GSK has approached CRUK on collaborating with their clinical trials registry: discussions are ongoing.

Katherine Littler updated the group on the Wellcome Trust's recent activities:

- Aside from the work on clinical trials, the safe havens work and patient data already discussed, the Wellcome Trust together with the MRC are due to release their new policy and framework on health-related findings from research early in 2014.

7. Incentives paper

David Carr introduced the EAGDA draft working paper on incentives for data sharing (circulated to the group in advance). The paper included the results of a web survey of researchers and data managers, several in-depth interviews with senior figures in universities and other stakeholder organisations, and a focus group with invited experts.

Key points from the paper were:

- There is a lack of support, rewards and recognition for those who generate and wish to share high-quality datasets. Costs to researchers in terms of time, financial costs and detraction from other research endeavours entail that data sharing is frequently not considered a priority.
- The implementation of data management and sharing plans is often not followed up by funders to ensure compliance, and they are not resourced adequately.
- Technical skills and resources for data sharing are often not prioritised in institutions and data managers are accorded a low status and few career progression opportunities.
- There is little, if any, recognition of data outputs in key assessment processes such as HEFCE's Research Excellence Framework, funding decisions and academic promotion.

The group discussed the paper and raised several issues for consideration:

- Time and concerns about protecting privacy are significant issues for researchers' data sharing efforts, and it may be worth further exploring these issues with early career researchers who were poorly represented in the survey.
- Many of the issues raised have been known for the past decade or so: it is worth considering why there has not been more progress towards addressing them over time.
- Three issues in particular affect researchers' data sharing:
    - The ability to share is contingent on adequate infrastructure and funding;
    - The desire to share requires evidence of its benefits;
    - The imperative to share must be backed up by sanctions for not sharing. These can have a significant impact on behaviour.
- Funders need to be prepared to invest in data sharing: this point should be made prominent in the paper.
- Review Committees do not usually have appropriate expertise for reviewing data sharing plans.
- Separate policies may be needed for existing datasets and for new studies, as existing studies may not have consents in place to allow sharing.
- Social sciences and genomics are both advanced in data sharing: this may be because they both have centralised infrastructures, which may provide a lesson in considering how medical fields can move forward with improving data sharing.
- EAGDA should advise funders to advocate for HEFCE to clearly and publicly recognise the contributions to scientific knowledge and research made by sharing datasets in the post-2014 REF (the relevant recommendation in the draft report should be strengthened). Data sharing is a legitimate form of academic activity that ought to be recognised and credited.
- Sanctions should be considered as an important tool for ensuring compliance with data sharing policies, and a proportionate approach should be adopted to enforcing data sharing plans.
- EAGDA could provide a strong recommendation to funders to adjudicate data sharing plans and ensure they are adequately resourced, and to monitor adherence.
- There are no benchmarks for data sharing costs, nor an established monitoring process for what works. It was suggested that EAGDA explore ways to assess research projects in which data sharing has been successful, and attempt to develop benchmarks to establish the true costs of sharing data.
- Centralised repositories need to be well resourced in order to function, but the costs of building and maintaining infrastructure (e.g., the EGA) are often overlooked.
- It is important to ensure that good practice is highlighted in the paper, as much progress has been made by funders in building infrastructure and developing policies for data sharing.

ACTION: EAGDA members to provide any additional comments on the draft paper to the Secretariat. The redrafted paper to be circulated to EAGDA members for comment by mid-November. It will then be decided whether to expand on the incentives web survey to seek further early career researcher opinions. Options for dissemination include compiling a short published statement to funders on incentives for data sharing, and producing the report for publication. This will be determined in early 2014.

ACTION: The Secretariat to discuss with funders options for further work on the web survey, particularly with early career researchers.

8. Identifiability paper

Natalie Banner introduced the EAGDA working paper on identifiability (following-up from the Gymrek paper discussed at the last EAGDA meeting), which examined the likelihood that research participants could be re-identified from their anonymised genomic data in the public domain in the UK. It suggested that although the technical risk of re-identification was very low, it is likely to grow as datasets become more complex, methods of analysis more sophisticated and datasets from different sources are linked together.

Key points from the paper were:

- It may be technically possible to re-identify individuals under specific sets of circumstances. However, it is important to note the distinction between being able to isolate the genetic characteristics of an individual ('individuation'), and linking this information to a name, address, location etc. ('identification)'.
- It is important to consider the motivation, means and opportunity of a data intruder attempting to re-identify individuals from genomic data.
- The boundary between what counts as 'anonymised' and 'identifiable' data may shift in light of the particular context in which the data are used and the datasets they may be linked with.
- Further research is needed into the actual risks and harms posed by the use and of anonymised data that could, in theory, be used in an attempt to re-identify individuals.
- With the exception of the ESRC, funders do not specify that sanctions will be imposed on data users who attempt to re-identify individuals from the shared data they have access to.
- Studies' consents should reflect that although the risk of re-identification is small, it does exist and that there may be future unanticipated risks despite anonymisation of their data.
- There is a need for public dialogue in this area, particularly with regard to the use of recreational genomics sites.

The group discussed the paper and raised several points for discussion and the further development of this work:

- There is currently significant public concern over access to, use and misuse of personal data by state and commercial organisations. This context needs to be addressed when considering the likelihood of data being misused and public trust in research uses of data.
- It was acknowledged that the likelihood of re-identification is low, but that new methods for re-identification are very likely to be developed in the near future. Clarity on sanctions against re-identification is therefore urgently needed.
- There is a substantial amount of public engagement work being conducted by the ONS and ESRC on the use of administrative data.
- Public work conducted thus far indicates that people are deeply suspicious of the use of personal data, but that clarifying that there are severe penalties and sanctions for

breaches in regulations or agreements for data sharing and use is a powerful tool for helping to build trust.
- The 1000 Genomes project provides a good example of a clear, simple and honest description of the risks of re-identification in its consent processes. This could provide a model for UK studies.
- There is a need for authoritative, practical guidance on how different data types might contribute to increasing the risk of re-identification. Empirical research would be needed to establish any quantitative metric or set of measures.
- It will be necessary to include in the paper discussion of the unauthorised use of data, given the current situation in which governments and other authorities have been found to be covertly using, processing and linking data.
- There was support for producing a statement to funders as matter of priority, highlighting the current and emerging issues with identifiability and making high-level recommendations for studies and funders on data governance and participant consent.

ACTION: The Secretariat to re-contact the ICO to establish whether the use of re-identified data (where the processing occurs outside the UK) could potentially be in breach of data protection principles and subject to penalties.

ACTION: The Secretariat to redraft the interim statement on identifiability for the funders.

ACTION: The Secretariat to explore options for joint work with NCOB on establishing the actual risks and harms arising from security breaches and infringements of privacy in the use of biological data, and the potential benefits from using these data.

ACTION: The paper to be expanded to include further details on sanctions, both legal and those that could be imposed by funders and institutions and to be brought back to the next EAGDA meeting for final sign-off.

9. Data Access Committees (DACs) paper

Natalie Banner introduced the EAGDA working paper on DACs, which broadly covers the governance of information release processes for relevant UK studies funded by the EAGDA funders. The paper included an analysis of DACs and other governance mechanisms, and the results of a survey of data users publicised in an August edition of *Nature*, accompanying a World View commentary by Martin Bobrow on the importance of developing good governance for data sharing.

Key points from the paper were:
- Lack of transparency both on datasets and the processes for accessing data is the key issue for many studies.
- Funders' policies specify the need for data sharing plans to be in place, but it appears that these policies are not being fully complied with.
- There is wide variability in the accesses processes used by different studies, across disciplines (with good reason). Many studies use generic steering or management committees to make access decisions, rather than having a dedicated access committee or group.
- The role of study PIs in making decisions about data access varies significantly.

- Users across different disciplines experience similar barriers in finding, accessing, using and linking data with lack of information, inconsistent labelling and metadata, and complexity of access procedures cited as particular difficulties.

The group discussed the paper and raised several points for discussion for the development of the paper:

- The distinction between discoverability and accessibility of datasets needs to be clearly drawn, as there are important differences in the challenges to each.
- It was agreed by the group that any efforts to standardise access procedures or create interoperable standards or descriptions would require international collaboration. It could be useful to liaise with the Data Access Control Office for ICGC on this issue.
- The MRC is building a data sharing gateway site to provide high level information on the data gathered by all of its cohort studies, including metadata. An on-going initiative with ESRC, CLOSER, is also aiming to create a universal search platform for data on nine co-funded cohorts.
- Efforts to develop interoperable standards can draw lessons from the creation of the world wide web: it is important that data formatting is machine readable and allows for a quick, lightweight search function allowing details such as study name, PI details, variables and data types to be listed (as opposed to requiring a centralised data repository).
- A 'toolkit' for studies setting up data access mechanisms would be a useful output from EAGDA, outlining guidance or best practice for when a DAC is needed, how they ought to operate (e.g., using a flowchart) and highlighting key points about data governance for studies to be aware of.
- From the user survey, it was noted that a high proportion (40%) of users did not return improvements such as cleaning or labelling data to the data producers. This ought to be an obligation for shared data users but enforcement was recognised as a problem. DACs could have a role in following up the use of shared data and imposing sanctions if needed.
- It was suggested that more emphasis could be made in the DACs paper on the possibilities for developing centralised DACs, which would enable studies to seed their data access responsibilities to an independent committee, staffed by experts and run efficiently. The P3G consortium in conjunction with the Centre for Genomics and Policy at McGill has very recently launched the IPAC, (International Policy interoperability and data Access Clearinghouse) to serve this purpose[1].
- In addition to EAGDA advising how funders should approach data access, a systematic consideration is needed of what institutions, commercial organisations, DACs themselves, study leaders, professional associations and journals need to do on the issue of data access.

ACTION: The Secretariat to look into ways to contact the relevant study PIs to follow up on the initial DACs analysis and attempt to evaluate when and why funder policies are not being complied with.

[1] http://www.p3g.org/news/official-launch-new-ipac-resource

ACTION: A fuller report to be written up for the next EAGDA meeting, with recommendations to funders. The feasibility of developing a 'toolkit' for studies to be explored.

ACTION: The Secretariat to research the P3G IPAC as a possible model for creating a centralised DAC for multiple different types of study.

10. AOB: EU Data Protection Regulation

Katherine Littler updated the group on developments with the EU Data Protection Regulation, outlining specific areas of concern for EAGDA:

- The LIBE committee is due to vote on 21 October (although this may be put back). It is at present unclear whether the Regulation will be passed before the next European elections in May 2014, at which point the draft may be rolled forward into the next Parliament, or reissued.
- Of particular concern for scientific research are the amendments tabled for Articles 81 and 83, which require that specific consent is needed for any use of any identifiable personal data.
- These amendments would, if passed, be extremely detrimental to scientific research involving human participants, and to clinical practice, as individual health records are often needed to be used.

The group discussed possible actions:

- EAGDA could compose a letter to the Department of Justice expressing concerns over the amendments and the likely detrimental impact not only on research, but also patient care and the sustainability of the healthcare system. This is particularly relevant given the plans for the 100,000 Genomes project.
- EAGDA could draft a list of specific arguments on how the Regulation could damage research and ultimately patient care, as a briefing for members and their associates.

ACTION: The Secretariat to keep a watching brief on developments in the European Parliament on this issue, and to continue working with colleagues and other funders to advocate for proportionate regulation of data use and sharing. Specific actions from EAGDA will depend on the outcome of the upcoming LIBE committee vote in Europe.

11. Next meeting date: 10 March 2014 (tbc)

Close