Safe Sharing of Research Data: The role of legal agreements when anonymising, 25th April 2019 – Summary

Number of delegates: approx. 120 from diverse teams within multiple UK universities (e.g. contracts, legal, research governance, information governance, Data Protection Officers, data repositories, trials units, etc)

Key messages from Rachel Smith, MRC Regulatory Support Centre:

- Identifiability is a continuum. The law however is binary. We therefore need to understand when we are working within the scope of the law or not.
- Identifiability is determined by both the content of the information (what direct or indirect identifiers are contained) and the context in which the information is viewed (e.g. what other information do you have access to?).
- Information can be considered anonymous once the content and context have been controlled to the extent that identification is no longer reasonably likely.
- When sharing, both the common law of confidentiality and data protection must be managed.
- Identifiable information can be shared. In a research context you would usually need to manage the disclosure of confidential information through consent[1], and comply with data protection law (if the information also falls under the definition of Personal Data).
- Re-identification from rich, individual-level data can occur, unless the content of the dataset and the sharing context are adequately controlled to prevent jigsaw identification.
- NHS England (Rachel Merrett) is working on a project to harmonise terminology.

Key messages from Vicky Cetinkaya, Information Commissioner's Office:

- GDPR defines Personal Data. It doesn't define anonymised information but describes it in Recital 26 (see slides).
- Pseudonymised data held within one organisation (i.e. where that organisation holds both the pseudonymised dataset and the cipher or key) is classed as Personal Data.
- When sharing data between organisations: if the recipient organisation is not reasonably likely to identify individuals, (i.e. has no reasonably available means to identify individuals) the recipient organisation may not be processing Personal Data.
- Rendering data no longer Personal need only be considered when an organisation cannot comply with the requirements of data protection.
- Further work is being carried out to clarify the relationship between data protection and common law of confidentiality with respect to anonymisation. There are likely to be circumstances when data is considered Personal but not confidential.

Key messages for Kerina Jones, Swansea University, Genomic Data Integration (GeDI) project:

- Genetic data covers a wide spectrum: from presence or absence of a trait to full genomic sequence.
- Not all genetic data is Personal Data: unique is not equivalent to identifiable.
- A risk management approach to sharing based on a traffic light system was discussed (see slide 9).

[1] Consent in this context is not referring to GDPR consent (as defined in the Regulation), rather consent that is sufficient to manage participants' expectations with respect to the common law of confidentiality.

Key messages from Alastair Nicholson, Health Research Authority:

- Alastair discussed the role of model agreements to enable data sharing for research.
- These agreements have been developed to support the sharing of data between NHS organisations and between a University and the NHS.
- You can find templates for the UK-wide model Non-Commercial Agreement (mNCA) and Commercial Model Clinical Trial Agreements (mCTA and CRO-mCTA) on IRAS.
- The HRA are keen to work with others to produce appropriate and relevant wording in agreements.

Key messages from workshop 1

The risks of not sharing data were discussed and the overwhelming conclusion was that we have to share data to support research. Some of the risks of not sharing were:

- Research is collaborative and data sharing is central to effective delivery (e.g. adequate sample sizes, secondary analyses, reproducibility, etc).
- Participant expectations cannot be realised without sharing.
- Research would not be cost effective, we would ultimately see a demise in research activity, number of research organisations, and research skills and capacity.
- There would be no progress in health care and hence no societal benefit.

We have to share.

Key messages from workshop 2

The risks involved in sharing data for research were discussed in workshop 2. These can be summarised as follows:

- **The researcher and what they propose to share**
  - Researcher track record: any previous breaches, lack of understanding of requirements.
  - Larger, richer datasets (both in terms of the number of data fields and the number of participants) posed greater risk of potential re-identification.
  - Rare occurrences pose higher risk.
  - Inclusion of 'special categories of personal data'. Within which it was recognised that some data may be more sensitive and thus have greater impact (e.g. flu status versus HIV status).

- **The recipient organisation and what they intend to do with the data**
  - Commercial collaborators, in particular small start-up companies were considered riskier than large pharmaceutical companies with established IG systems.
  - International collaborations are higher risk.
  - Research involving large consortia / numerous partners are higher risk.
  - Complexity of projects e.g. high number of data linkages increases risk.

- **What participants have been told about the sharing of their data**
  - Whether there is consent[2] for the sharing and what transparency information says.

[2] Again, consent in this context is not referring to GDPR consent (as defined in the Regulation), rather consent that is sufficient to manage participants' expectations with respect to the common law of confidentiality.

Key messages from workshops 3 and 4

Workshops 3 and 4 focussed on managing the risks of sharing covering:

- assurances that the content is adequately controlled by the research team; and
- how the context can be controlled.

- General agreement that data sharing should be considered on a case by case basis with a view to categorising and applying appropriate controls (one size does not fit all). The triage system will require case by case scrutiny, however the risk management strategies employed could be developed to reflect risk categories.
- Difficult to develop a detailed data sharing risk assessment tool (depends on the University's research portfolio and systems – what's considered high-risk by some may be daily business for others, and this may change over time). However, a general framework to support decision-making may be possible.
- It was agreed that only researchers have the expertise to effectively control the content of the data. Governance staff in some organisations intuitively trust researchers do this appropriately.
- There are a number of assurances that the researcher can provide governance staff to demonstrate trustworthiness (e.g. good track record, their approach and research data flows hang together, they understand anonymisation, they engage with governance processes; they use a trusted service like a CTU, statistician, safe haven (see below), central repository). Having discussions and building relationships between governance functions and researchers helped to engender trust.
- Some universities use existing governance processes as far as possible to provide assurances (e.g. Data Management Plans, Data Protection Impact Assessments, etc.). However, more could be done on bottoming out exactly what assurance these existing processes bring and when/how other departments should take account of them.
- Agreements were discussed as a primary means for controlling the context of data sharing and a number of clauses or phrases to include were highlighted. Further work in developing model agreements where the NHS is involved is being carried out by the HRA; and for university to university transfers by the Russell Group led by Nottingham.
- The idea of controlling context through 'safe havens' aka 'trusted research environments', was discussed. It was clear that these are quite diverse, and mean different things to people. Further work is needed to define what they are and what their role is, or could be, in the future.
- Acknowledgement that context controls would need to be reviewed and revised as research evolves and more relevant information is in the public domain. In some fields this is more rapid than others, e.g. genetics.

Next Steps

As a result of the workshop we have identified the following next steps. We will work with stakeholders to identify who is best placed to take each forward:

- **Clarifying the relationship** between data protection and common law of confidentiality with respect to anonymisation.
- **Risk-based triage tool** – Develop a decision-making framework to support risk assessment and use of appropriate controls in different risk categories, perhaps based on a traffic light system.
- **Safe havens / trusted research environments** – There needs to be a common definition/set of standards, agreement on when these should be used and their benefits. All of which needs to be communicated to researchers and those involved in governance (including information governance and those who have oversight of contracts).

- **Agreements** – A number of stakeholders (i.e. HRA, Russell Group and/or Brunswick Group) are already working on the development of model agreements. Are there opportunities to join up this work and learn from each other?
- **International research** – The difficulties of sharing data internationally for research was raised at a number of tables. Scoping out international aspects further, e.g. difference in definition of anonymous?
- **Good corporate governance** – What does good governance look like? e.g. promoting cross departmental relationships, taking account of others' processes, reducing duplication and burden.
- **Public engagement** – There's a need to educate the public about research generally (some stakeholders might be well placed to lead or feed into this work e.g. Understanding Patient Data).