Turning policy into practice

Barriers to data sharing, as seen by the people who make the data

Neil Walker, JDRF/Wellcome Trust DIL

20th August 2015

OK, title a bit hubristic – groups differ – but:

Data sharing is onerous, especially if you haven't planned it well.

Assume we're not discussing the big infrastructural projects at e.g. the Wellcome Trust Sanger Institute.

Assume also that the Data Management Plan lodged on grant submission lacked implementation detail . . .

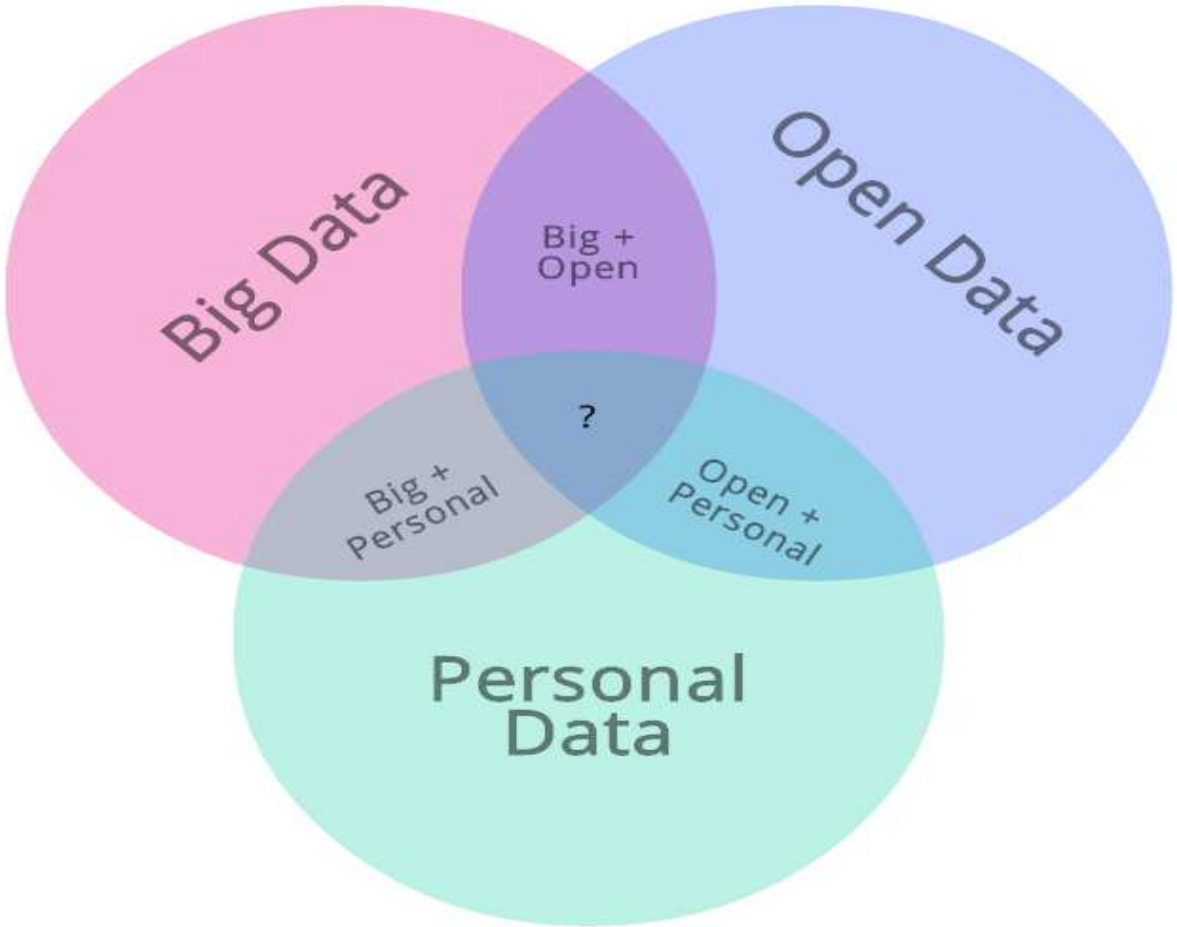
Who handles the data?

As genotyping and sequencing become cheaper, are offered as a service, genetic data may reside with small clinical research groups with no data management expertise or training.

The person with the data may lack confidence in, and support for, data sharing . . .

What are the sources of confusion, and where does the support come from?

1. Confusion between Open, Personal and Big Data



The Open Data Institute: `https://github.com/theodi/data-definitions`

1. Confusion between Open, Personal and Big Data

Genetics is Big and Personal.

Shared data is not the same thing as Open Data.

Even aggregate data need not be Open Data.

e.g. Immunobase: `https://www.immunobase.org`

2. Confusion over policy exemptions

If genetic data is personal, can it be shared at all?

Considered an exemption in EPSRC and BBSRC policies . . .

2. Confusion over policy exemptions

MRC say personal data can be used in research, and shared if consented.

> *"Existing data sets can be shared with other researchers provided this is not inconsistent with what participants were told about how the data would be used."*

Section 7.2.1, Personal Information in Medical Research:
`http://www.mrc.ac.uk/documents/pdf/personal-information-in-medical-research/`

3. Confusion over consent

However, boilerplate *"Participant Information Sheet"* template from MRC/NHS Health Research Authority asks:

- How will my information be kept confidential?
- What will happen to the results of this study?

Standard answers (much truncated):

- All information collected about you as a result of your participation in the study will be kept strictly confidential. Your personal and medical information will be kept in a secured file and be treated in the strictest confidence.
- The results of the study will be anonymous and you will not be able to be identified from any of the data produced.

`http://www.hra-decisiontools.org.uk/consent/examples.html`

3. Confusion over consent

UKDA say (in answer to hypothetical questions):

Q. If I ask my respondents for consent to share their data then they will not agree to participate in the study.
A. Don't assume that participants will not participate because data sharing is discussed. Talk to them – they may be less reluctant than you might think, or less concerned over data sharing!

Q. I am doing highly sensitive research. I cannot possibly make my data available for others to see.
A. The first thing is to ask respondents and see if you can get consent for sharing in the first instance. Anonymisation procedures can help to protect identifying information [or] consider controlling access to the data.

Exercise 2: `http://data-archive.ac.uk/media/431116/trainingresourcespack.pdf`

4. Confusion over anonymity

Without consent, researchers try (and fail) to anonymise data.

> *"There is no evidence that de-identification works either in theory or in practice and attempts to quantify its efficacy are unscientific and promote a false sense of security by assuming unrealistic, artificially constrained models of what an adversary might do."*

Narayanan A & Felten EW (2014) No silver bullet: De-identification still doesn't work
`http://randomwalker.info/publications/no-silver-bullet-de-identification.pdf`

4. Confusion over anonymity

ESRC use risk assessment model.

Almost nothing too confidential to share, or Open Data – everything on spectrum in between.

Impact Level	Data type	Access method	Notes/examples
0	Open Data	freely available	*no individual-level data is released under this method*
1	non-disclosive	on registration	*unlinked survey, genetic data*
2	potentially disclosive	on application	*linked genotype/phenotype*
3	disclosive	in a secure setting	*health, tax, school records; detailed geographies*
4	confidential	not available	*census data, archived and released after 100 years*

NB: requires data governance structures.

5. Confusion over what to share

Perfect is the enemy of the good. Is there a minimum dataset?

1. raw data (in terms of samples)
2. processed data (in terms of samples)
3. a list of sample exclusions (technical failures etc)
4. a sample-to-subject lookup
5. clinical phenotypes (by subject)
6. subject exclusions (if any)
7. a README to describe data provenance and methods

NB: need external work on variant relevance and ontologies.

6. Confusion over where to share data

- MRC has no repository
- Wellcome Trust's clinical repository is in development
- Dryad and Zenodo do not allow personal data
- figshare is focused on Open Data – private sharing mode currently for pre-publication collaboration?
- Wellcome Trust sponsors EBI's access-controlled repositories, very suitable for genetics
- alternative is to post metadata on institutional website, and issue data on request.

1. How to get institutional support?

Really EAGDA territory, but . . .

Following success of Open Access mandate, observable that mandating data sharing engages institutional support (EPSRC, BBSRC): requesting but not auditing data sharing (MRC, Wellcome Trust) does not.

Lightest convincing level of audit?

"EPSRC will investigate any complaints about research data not being managed in line with EPSRC expectations."

2. How to get PI support?

I have PI support. However, there is always something more pressing than a “nice to have” data sharing project at the end of a study.

More general PI support for data sharing will follow if:

- funders insist on it
- it is planned for in advance
- journals demand it – and check it
- data sharing gains credit in RAE, grant applications

Could do the last via citations? Data sharing increases the number.

Piwowar HA & Vision TJ (2013) Data reuse and the open data citation advantage

`http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3792178/`

Conclusion

IMO, genetic data sharing is in good health, but that doesn't necessarily make the shared data useful, *and* it is threatened by spread of genotyping technology to those unfamiliar with the habit and practice of data sharing.

Education, plus some sticks and carrots to get institutional buy-in, would go a long way.

Extra slide: JDRF/Wellcome Trust Diabetes and Inflammation Laboratory Data Management Plan

"We intend to be as open with trial/study data when analysed and published as is consistent with participants' informed consent. In practice, that means we publish anonymous aggregate data through peer-reviewed journals and conference presentations; and (on publication) archive anonymised individual-level data in community-endorsed access-controlled public repositories (e.g. dbGap, EGA) where available. Where no suitable repository can be identified, anonymised, individual-level data will be available from our website, to named, relevant, bona fide researchers, on completion of a Data Access Agreement. Data will be discoverable both through accession records in peer-reviewed publications, and through lodging metadata with our institutional repository."