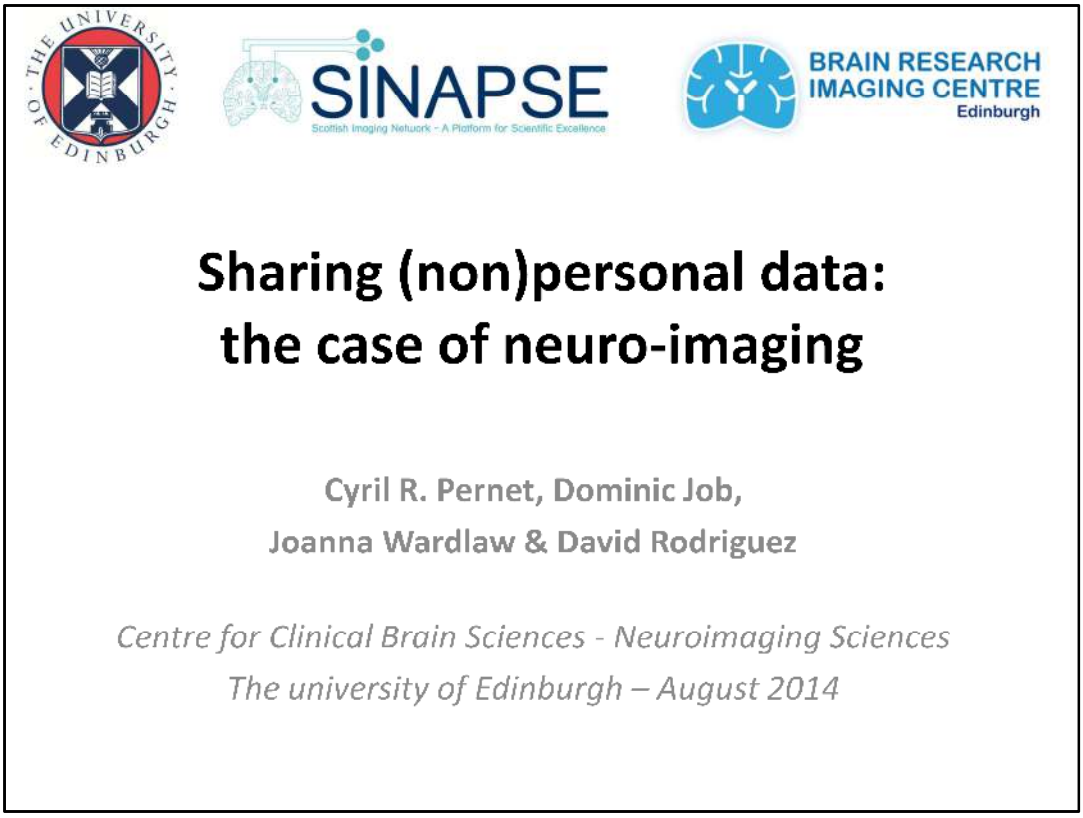


Dealing with Data Conference – University of Edinburgh 26th August 2014

Why sharing data

New rules & Better science

The rules have changed and pretty much any research data should be shared
1 - The OECD describes data as a public good that should be made available
2 – RCUK says data should be preserved and accessible for 10 years +
3 – All major funders require data to be available

Sharing data also makes science better by
1 - accelerating progress in our fundamental understanding of the topic addressed by the data
2 - improving publication and data quality
3 - making research reproducible
4 - fostering research and advances in practices
5 - reducing the cost of research and increases the return on current research investments

What are Personal data?

Personal data are data which relate to a living individual who can be identified –
(a) from those data, or
(b) from those data and other information which is in the possession of, or is likely to come into the possession of, the data controller.

These are things such as an individual's personnel file, their medical records or home phone number.

A further tier of personal data is defined in the Data Protection Act: sensitive personal data. Sensitive personal data includes data relating to a person's race, sexuality, health, criminal record or affiliations (such as political persuasion or trade union membership).

Data protection act

Personal data must be

1. Processed fairly and lawfully
2. *Processed for special proposes*
3. Adequate, relevant and not excessive
4. Accurate and kept up to date
5. Not kept for longer than necessary
6. Processed in accordance with the rights of data subjects
7. Protected by appropriate security (practical and organizational)
8. Not transferred outside the EEA without adequate protection

Resources:
http://www.legislation.gov.uk/ukpga/1998/29/contents
http://ico.org.uk/for_organisations/data_protection/topic_guides/data_sharing

1 to 7 deals with how, as researcher, you should collect and handle personal data – but item 2: sharing data allow secondary use (for research this is allowed)
Item 8 deals with sharing.

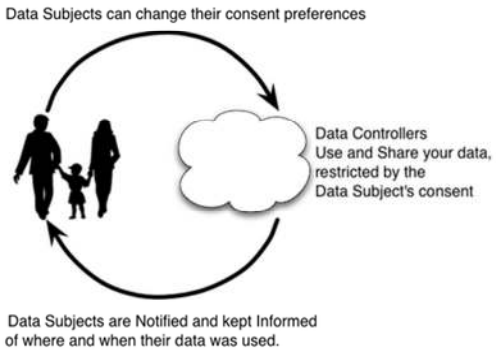
Sharing (non)personal data

- From now on I will talk about sharing personal data securely via data bases and repositories

- information consent forms
- Withdrawals
- De-identification
- Cross-linkage

Informed consent forms

- Agreement to the further use and sharing of the data beyond the original purposes of the study - Is a broad consent, an informed consent?
- Emergence of 'dynamic consent'



For most research information sheet a given to participants and consent forms obtained – usually these forms are treated as a one-off event - Laurie et al. 2013

To accommodate the need for data sharing these forms now include an agreement for further use and sharing ; the question raised is thus of how much information is given to be an informed consent – there has been much debate about broad consent with the emergence of biobanks and future consent should take example of what has been used to ensure participants understand well to what they are agreeing (Hansson et al. 2006, Sheehan et al. 2011, Steinsbekk et al. 2013).

It has been argued that consent should be seen as a continuing relational process that will evolve over time – some project now have dynamic consents in which participants can be contacted to agree for every new research conducted on their data.

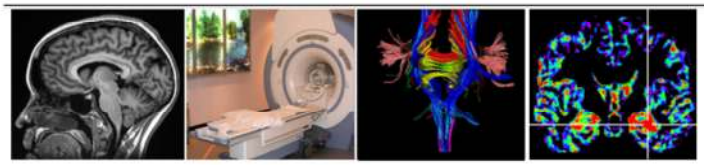
Withdrawal

Subjects have the right to withdraw their data from a study at any time. Once the de-identified data are on a public database it becomes almost impossible to remove them– the information consent form must therefore make provision that it may not be possible to remove data once it has been shared.

For research which has the potential to generate important clinical information about individual participants, arrangements should be made to allow some party (e.g. data-base manager), to keep a key to re-identifying data sources should the need arise.

De-identification

- BRAINS database : structural data + metadata



BRAINS: Brain Images of Normal Subjects

Brain Images of Normal Subjects bank is being developed with more than 1000 normal subjects from across the lifespan. It is collating images, and associated information (metadata) about health (e.g. blood pressure) already collected from people participating in research projects throughout Scotland. Many of these studies include detailed information from across the whole lifecourse, including socioeconomic status, current and previous health, medication use and cognitive ability tests. We are initially focussing on collecting data from studies at the extremes of life (old age and pre- and neo-natal) where there is most variability in brain structure, but we aim to expand the bank to include subjects of all ages.

The bank may be expanded in future to include subjects from other geographical locations, and patients with a range of neurological disorders, e.g. Alzheimer's disease, stroke, and schizophrenia.

The images have been collected in imaging centres across Scotland and are in a range of magnetic resonance (MR) sequences, including T1, T2, T2*, and fluid attenuated inversion recovery (FLAIR). When BRAINS is released these will be searchable by a wide range of metadata, e.g. blood pressure, age, MMSE.

BRAINS atlases are based on calculated distributions of brain structure rather than parametric estimates. These will be used to support image analysis research and clinical reporting of brain images.

De-identification is the process used to prevent a person's identity from being connected with information (http://en.wikipedia.org/wiki/De-identification). When sharing, data descriptors should not contain personal information and the data themselves should be made unidentifiable.

As a concrete example let's look at the data base we are developing the Brain Research Imaging Centre – the data are structural brain images of healthy subjects of all ages + metadata about age, health, medication, etc

De-identification: imaging data

- Image header contain patient information

Tag	VR	Description
(0008,1120)	SQ	ReferencedPatientSequence
(0010,0010)	PN	PatientName
(0010,0020)	LO	PatientID
(0010,0021)	LO	IssuerOfPatientID
(0010,0022)	CS	TypeOfPatientID
(0010,0024)	SQ	IssuerOfPatientIDQualifiersSequence
(0010,0030)	DA	PatientBirthDate
(0010,0032)	TM	PatientBirthTime
(0010,0040)	CS	PatientSex
(0010,0050)	SQ	PatientInsurancePlanCodeSequence
(0010,0101)	SQ	PatientPrimaryLanguageCodeSequence
(0010,0102)	SQ	PatientPrimaryLanguageModifierCodeSequence

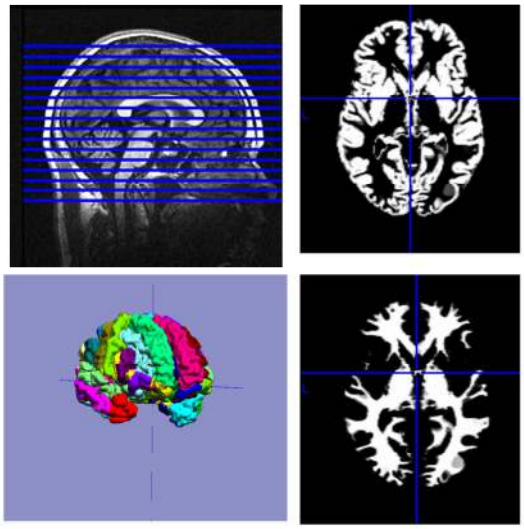
De-identification consists in keeping all information but personal data – each medical image (DICOM) image come with descriptors - here we'll have to remove the patient name, and ID, DoB – and substitute name/ID by a code.

For research which has the potential to generate important clinical information about individual participants, linked anonymised data should be used allowing some party (e.g. data-base manager), to keep a key to re-identifying data sources should the need arise. If no potential clinical value is foreseen, unlinked anonymised data can be used.

Source: http://www.dicomtags.com/

De-identification: imaging data

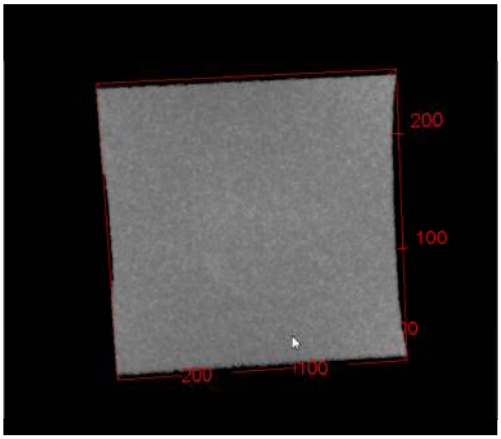
- Image reconstruction



Researchers accessing such data are interested in looking at the brain, extracting various features like gray matter mater volume, the size of some specific region etc

De-identification: imaging data

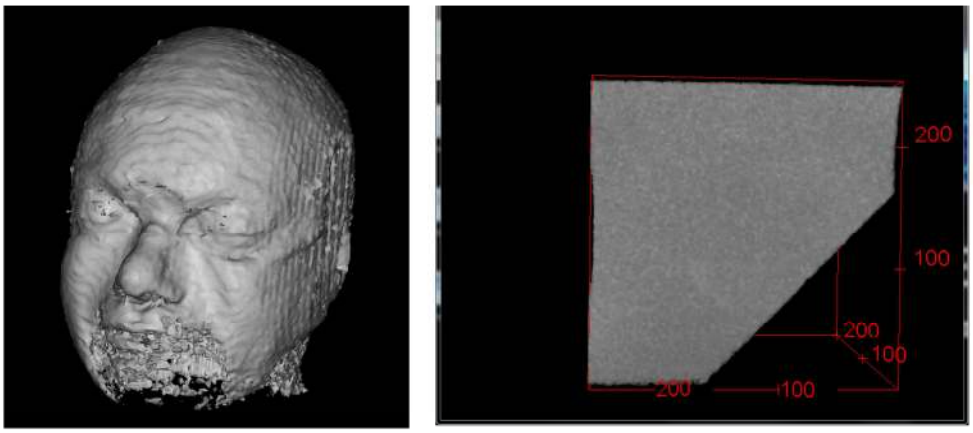
- Image reconstruction



But MRI has such a good spatial resolution that one can recognize individuals -- here we start the raw data and simply look for the right threshold in the data (video) ; and this guy appear – maybe some of you know him, he will give a talk later on this morning.

De-identification: imaging data

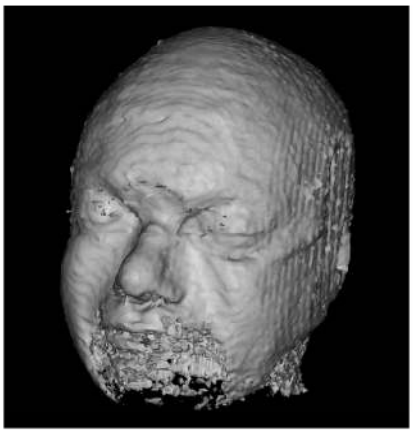
- Image reconstruction

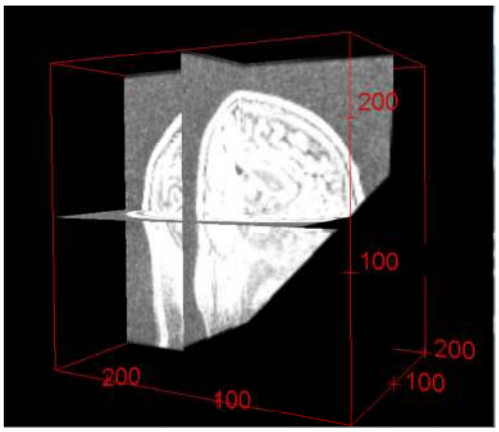


Because we have to share data, we cannot provide these data as they are – even when the metadata have been cleaned.
One simple solution can be to find the right angle in the data to remove the face .. (video)

De-identification: imaging data

- Image reconstruction





Note that is not just the face that matters – personal data are data which allow individuals to be identified, which can be the case using dental record or ear morphology – so our data must not contain this information otherwise, by crossing data bases, it becomes possible to identify people

The issue of re-identification is real ; by crossing information from different data bases it has been possible to identify participants in the 1000 Genomes Project by combining publicly available demographic information from the American census and public information from peoplefinder.com with anonymised genomic datasets (Gymrek et al., 2013). Last March, the Wellcome Trust. Cancer Research UK, the MRC and ESRC have outlined steps to take to de-indentify data properly – this concerns genetics but also in general medical, epidemiological and social sciences.

A major point I want to stress here is the code to use to de-identify data has to be generated using dedicated methods – for instance The NYC Taxi *geolocation* data base was made available but used a 'one way hash without a salt' (ie a simply cryptographic function without random element) which allowed to re-identify specific taxi drivers. Of course, if one share unlinked data – a simple 1, 2, 3, 4, 5 could do the trick.

Source http://www.wellcome.ac.uk/News/Media-office/Press-releases/2014/WTP055974.htm
http://www.futureofprivacy.org/2014/07/24/de-identification-a-critical-debate/

Conclusions

- Be aware of the principles in the data protection act
- Share your data !

- But use 'updated' consent forms based on how you are going to share data
- Be careful with the de-identification method used – if there is no need to use linked data, don't ! (random numbers are great)
- Make sure nothing in the data allow re-identification