Keeping Anonymous Data Anonymous

Rupert Major, Department of Health Sciences, University of Leicester, Leicester, UK.
Supervisors: Dr Laura Gray and Professor Nigel Brunskill

Problem

Large population studies may contain thousands or even millions of patients. We are developing 2 studies with a combined population of 60,000 patients with kidney disease. In this situation, it is not practical to gain the permission of each and every patient. We therefore use **anonymised data**.

This can potentially create 2 further problems that might allow **identification** of an individual in the study:

1. Information ('pseudoidentifiers'), such as date of birth or postcode, may allow the anonymised individual to be narrowed down to less than 100 individuals, combining more than one pseudoidentifier would allow **re-identification** of individuals.

2. Combining data from other data sources ('**linking**'), such as hospital records, could also allow identification of individuals.

1. Preventing Re-identification

Reducing how unique information is will reduce the chance of an individual being re-identified. For example:

- A postcode allows identification of the street somebody lives on.
- Date of birth is unique to ~1 in 20,000 individuals.

We can modify data before it is extracted to make it less unique though.

Original Information		Date of Birth	Postcode
People data unique to	Individually	60	6
	Combined	1	
Modified Information		**Age in Years**	**Postcode based Poverty Score**
People data unique to	Individually	1,200	5,000
	Combined	100	

Table 1: how modifying information can make it less unique and reduce the risk of re-identification.

By using age and poverty score instead the data has been changed from unique to 1 individual to ~100 individuals, thus reducing the risk of re-identification.

2. Linking to Other Datasets

Patients will often have data stored with their GP and in hospital. For research purposes access to all of this data is very useful but it presents a potential risk for re-identification.

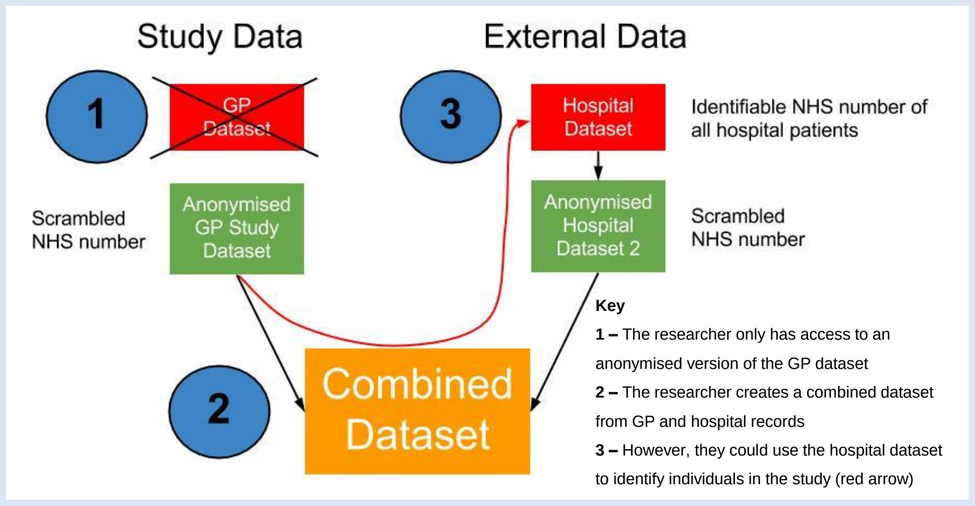

Figure 1 – how individuals in a study could be identified from another dataset.

So how can they be linked together without breaking anonymity?

Figure 2 shows the process to protect anonymity.

1. Datasets are separated so that one researcher will never have access to both.
2. Before data is extracted a unique anonymised identifier based on the individual's NHS is created. Researcher 1 therefore does not know the identity of the study's participants.
3. Researcher 2 (a hospital clinician) then performs the same process for the whole of the second dataset, including those not in the first dataset.
4. This data is sent to Researcher 1 and the scrambled NHS number is used to link the data. Unused data is then digitally destroyed.

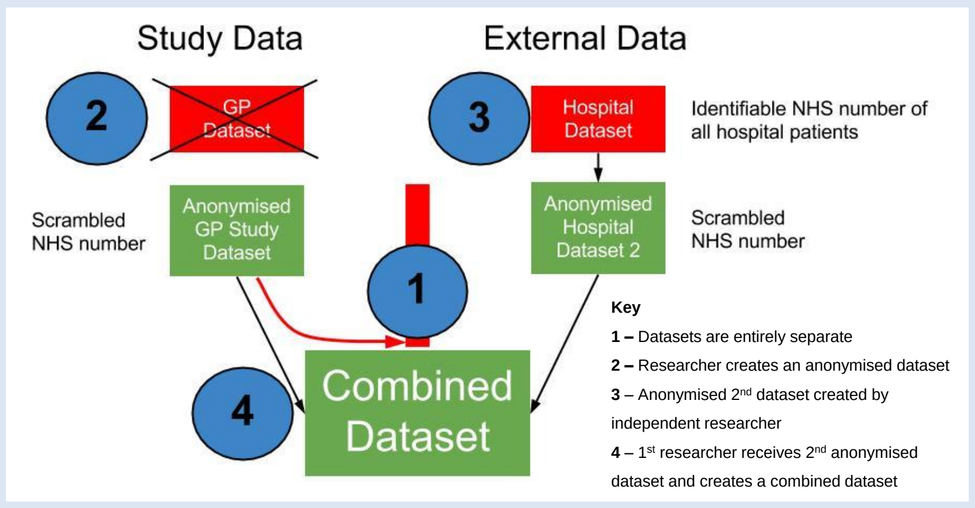

Figure 2 - modified process to prevent re-identification of individuals.

The 2 datasets are now linked without breaching the anonymity of individuals in the study.

Conclusions

In large studies of thousands or millions of people it is not practical to gain permission of every individual so anonymised data is used. However, despite being anonymised some information might allow identification of individuals from study data. Whilst certainty of anonymity can never be 100% guaranteed, we can use techniques to substantially reduce the risk of re-identification.

I would like to thank Kidney Research UK for generously funding my PhD.

rwlm2@le.ac.uk , Twitter: @KRUKPhD , Youtube: Rupert Major
Blog: https://www.kidneyresearchuk.org/blogs/rupert-major