Techniques for Data Linkage and Anonymisation

David J. Hand, Imperial College, London

23rd October 2014

Most things in life involve a balance. We balance short term gratification against long term return, risk against reward, cost against benefit, and privacy against openness. The Court of Justice of the EU noted that the right to data protection is not "an absolute right, but must be considered in relation to its function in society".

And our topic today is a perfect example of such an attempt to strike a balance.

It's clear that linking records from different sources can be hugely beneficial - to individuals and to society at large. A familiar and obvious example is the linking of diet and exercise to health. A more subtle one is linking data on housing characteristics and location with data on mortality and morbidity.

But it's also clear that leaking confidential information can be seriously detrimental, most obviously to individuals, but also to society at large. Speaking of the inadvertent leaking of medical information, MP George Mudie said "*The human cost to the patient whose identity and medical history are made public is potentially disastrous. Careers could be ended, jobs lost, insurance refused and relationships destroyed if sensitive medical facts are made public or used by private firms, other people or, indeed, the media*."

Now in many situations it's merely a question of finding a position for the fulcrum of the balance - deciding how much risk should be compensated by how much reward, or in our case, how the chance of leakage of confidential information is compensated for by the potential gains to individuals and society. In other situations, however, by astute thought, design, and planning, one can change the nature of the balance, so that reward can be increased without impacting risk, or at only a small increase in risk.

I think today's topic falls into this class. Our aim today is to discuss some of the initiatives which are going on in this area, the potential gains to be made by record linkage in a number of very diverse areas, while also showing tools for linking data as accurately as possible and controlling the associated risk.

One might characterise techniques for reducing risk in data linkage as functioning at at least three levels:

(i) technical - anonymisation, de-identification, trusted third party methods, etc
(ii) procedural - the use of secure rooms, the accreditation of researchers, etc
(iii) legal and regulatory - etc.

(i) Technical:

Technical aspects themselves occur at at least two levels.

(a) encryption methods. These are important when data are transmitted between computers, and can be important when data are stored.

(b) anonymisation and de-identification methods. For example, trusted third party strategies separate the linkage aspects from the content, so that the owners of neither database hold the linked records and so that no party holds linked records along with identifiers.

(ii) Procedural:

An example is secure rooms where analysis is undertaken, with no mobile phone signals, no data taken into or out of the room, no USB or other ports on the computers. And so on.

(iii) Legal and regulatory:

This may include the requirement that researchers have to undergo training and accreditation procedures, and that organisations involved in such exercises have to be accredited.

At a higher level, there will be laws which constrain the way organisations can use data, the requirement that they must respond to queries about the stored data and how it is used, and so on.

It has to be said that the legal environment is in a state of flux at present. The UK Cabinet Office has been considering a possible data sharing bill, looking at three aspects: fraud, error and debt; tailored public services for individuals; and research and statistics. At the EU level, a new Data Protection Regulation is being proposed which would severely restrict the use of personal data for scientific research purposes and which could have significant adverse impact on health and social research - and, of course, as a consequence, the effectiveness of public policy. Most recently, the UK Law Commission tabled a paper to Parliament recommending a full three year review of data protection law.

I've characterised things as a balance between the benefits and the risks arising from linking data. But I don't want to give the impression it's a simple choice. It's not a question of "if you link you gain the benefit and incur the risk, while if you don't link you forgo the benefit and avoid the risk". It's more subtle than that.

It can in fact be the case that "if you don't link you incur *other* risks, possibly worse than the ones you had hoped to avoid".

Here's a very topical example.

You will all be aware of the *care.data* intiative. The nature of the balance it sought to strike between public good and individual risk has been hotly debated. Indeed, for patients who are particularly concerned, it is possible to opt out - that is, you can choose not allow your GP data to be brought together with hospital data. The website of one of the privacy groups says "opting out will not affect the care you receive". But I don't think that's quite right. Anyone who chooses to opt out is degrading the database. Furthermore, they are not degrading it in a random way. Now, to this audience, I don't need to spell out the danger that brings. One of the most common causes of mistaken interpretation of a statistical analysis arises when the data are incomplete in some non-random way – when they are a

distorted representation of the population about which one hopes to make an inferential statement. When this happens, the conclusions drawn can be very misleading.

At an extreme, if people with certain characteristics all refused to allow their data to be included, then any care they received would be based on analysis of people who differed from them. If no-one with disease X allowed their data to be included, what could we say about the lifecourse, the treatment, and the prognosis of disease X?

So what this means is that anyone choosing to opt out is potentially invalidating the statistical conclusions, putting at risk not only the future care they might receive, but also the future care others might receive.

I'm exaggerating - statisticians are very clever at devising methods to alleviate such problems - but the danger is a real one.

At this meeting, as is appropriate for a programme held at the Isaac Newton Institute for the Mathematical Sciences, most of our attention will be on the technical - more mathematical aspects. But the first session will put things in context, giving a high level views of the two research councils most concerned with these issues.