Response to Nuffield Council on Bioethics consultation on the linking & use of biological and health data

Prof Jeremy Wyatt & Dr Hamish Fraser, eHealth Research Group, Leeds Institute of Health Sciences, University of Leeds

We welcome the Nuffield call for comments on this timely and important topic, agree with its scope and aims and look forward to reading the report's conclusions and recommendations in due course. Our responses follow.

Question 1: Special significance of biomedical data

Is it helpful – given the large and increasing amounts of behavioural data being collected and used in public and private sector organisations – to retitle this category of data "Data on human health and behaviour", or "Human biomedical and behavioural data" ?

Special characteristics of genomic data: it is impossible to fully anonymise a complete gene sequence or significant part of a gene sequence – thus explaining their use for forensic purposes. However, short sequences or abstractions of gene sequences such as mutations can be anonymous – unless there is only one person worldwide with such a mutation.

Question 2: New privacy issues

Public interest: this might take precedence over private interests when the interests or losses due to ignoring them are unlikely or small and the public benefits are likely to be large. One tool to help assess the likely public benefit might be the Expected Value of Perfect Information – Karl Claxton, York.

Actual harms we must avoid: these range from the rare but serious harms (eg. an organised crime syndicate unmasking the identity of a witness who has been given a new identity, with potentially fatal results) to common and aggravating harms (eg. identity theft with moderate financial loss). We need better evidence about the rate and severity of harms resulting, and propose setting up an open anonymous web based register of confirmed harms arising from privacy breaches to which the public and/or professionals would be encouraged to contribute. While there would inevitably be selective reporting, trends over time or across geographical areas could still yield useful insights into the type, severity and frequency of privacy breaches. Analogous situations arise in business and the financial sector where fraud and security breaches are usually not discussed publically to "protect the reputation" of the organization. Is there (should there be) a requirement that all health data security breaches or misuse cases are reported so that their true incidence and impact can be monitored?

Why does it matter if data are used in ways of which people are unaware: our qualitative study of public attitudes to the use of anonymised data for research revealed some surprising insights about how people still felt that anonymised data were "their" data, and wished to be consulted about its use and told about any study results. See:

Haddow G, Bruce A, Sathanandam S, Wyatt JC. 'Nothing is really safe': a focus group study on the processes of anonymizing and sharing of health data for research purposes.. J Eval Clin Pract. 2010 Jul 13.

Question 3: Impact of developments in data science etc.

Impact of availability of biomedical data on how research is funded: the Research Councils (ESRC & MRC) have between them spent over £120M in the last 2 years setting up the Farr Institute and networks of centres to investigate data applications in healthcare, medical bioinformatics and social sciences and the four Administrative Data Research Centres. As well as this funding, another reason for the shift towards biomedical research using big data is the cost and over-regulation of clinical research, and especially academic clinical trials.

What are the main barriers to development and innovation – in my opinion, too much of the work in this area is driven by the availability of data and of technology to manage it, and too little by clearly formulated research questions developed in consultation with policy makers and the public. In other words, too much eScience is technology led, not science or impact led.

Significant developments in linking and use of data to which Nuffield Council should pay attention:

- Text mining (the extraction of meaningful data from unstructured free text) – eg. by Google text mining emails to target adverts to Gmail users; the IBM Watson technology with its potential to scan and detect trends or predictive insights from social media, emails, corporate records etc.
- The development and use of virtual research environments pioneered by the ESRC data centre in University of East Anglia, in which data are never distributed, but only made available to accredited researchers in a controlled, locked down environment using eg. Citrix. All outputs of the analysis are held in the VRE until they have been scrutinised for identifiers or other material that might increase the chance of unintended disclosure
- The use of deliberate small data manipulations on a random subset of data records to deter potential attempts at re-identification – "Barnardisation". Sheila Bird, an MRC statistician in Cambridge is an expert on these techniques.
- Work of Latanya Sweeney at Harvard University on the risks of re-identification of personal data - including genetic profiles - based just on the demographics of individuals: http://dataprivacylab.org/projects/pgp/index.html
- The technique of using datamaps (http://thedatamap.org/index.html) to help people visualise how their data are used by organisations and understand the potential privacy implications – eg.

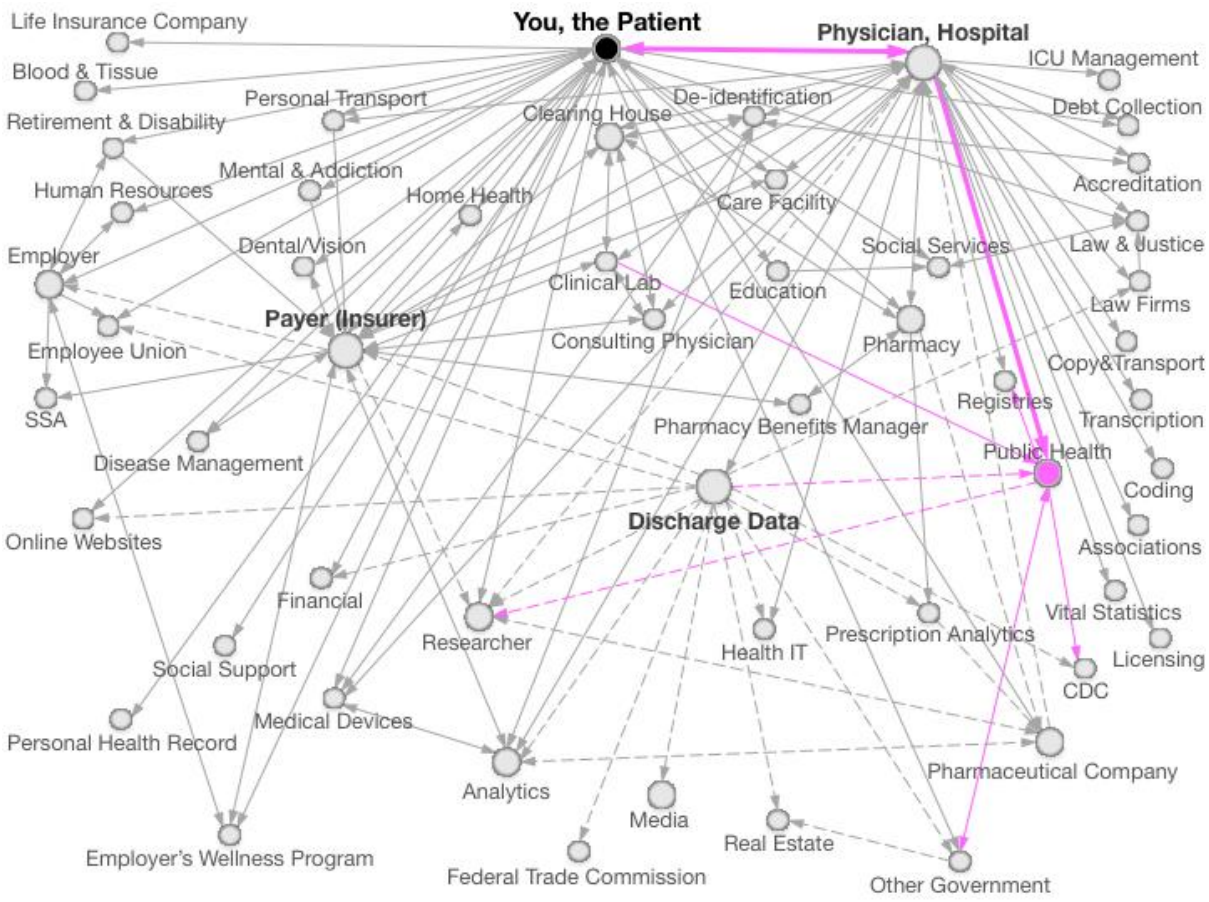

Figure: Datamap for a typical patient member of a US Health Maintenance Organisation, to help them appreciate the extent of data use and potential privacy implications – from http://thedatamap.org/index.html

Question 4: opportunities for & impacts of using linked data in research

Allowing access by others to data researchers have collected: yes, especially if the research is publically funded; but it seems reasonable to allow the original researchers a period of time – 6-12months – during which they get sole access to the data.

Research carried out by commercial firms: if participants were members of the public or users of a public service, there should be an obligation to publish within a reasonable timeframe, whether the results were positive, or negative / revealed harm.

Incorporation of public sector data into a private sector dataset / analysis: a helpful analogy may come from open source software. In most open source / FLOSS licences, if any open source code is incorporated into a new product, that product should also be published as open source. For data, if any of the data used in research originated in the public sector, then the analysis should be published openly.

Question 5: opportunities and impacts of data linkage in medical practice

Personal refusal to allow data usage for research even when that person benefits from a public service: maybe it is reasonable to require data usage unless the consequences of unintended disclosure cross a threshold for the likely damage that could result (see example of person on crown witness protection scheme). One analogy is notifiable diseases: here the public good of preventing an epidemic allows the communication and use of data without consent. This dates back many decades to when infectious disease epidemics were common and serious. However, we have other epidemics now that could arguably require data sharing without consent. According to Dame Onora O'Neill, the law is weak at addressing the needs of society & the public interest and much better developed at protecting individual interests. I'm not advocating a Maoist approach ("*There is no private data, all information belongs to the State – get over it*"), but we do need to redress the current rather libertarian focus on personal rights, ignoring the impact of this on public good. The recent debates around NSA surveillance have shown that despite concerns, most people recognise a need for some government surveillance to protect them from terrorism. This provides a useful analogy to the obligatory anonymised use of data to protect society from epidemics etc.

Question 6:using biomedical data outside of research and health care

Should individuals be able to profit from their data ? Tools like Mydex (in which people maintain and control access to their own data, and organisations "rent" it from them for a specific purpose and time period) will inevitably lead to this. While some ethicists might like to make the analogy between personal data and blood or organs for transplant (for which people cannot be paid – in the UK at least), there is a fallacy here: data can be copied indefinitely, but we cannot donate blood or organs an infinite time.

What are the ethical implications of using predictive analytic tools with biomedical data outside health care and research (e.g. in recruitment or workforce management)? The complex nature of IT systems used by companies and organization to determine insurance rates, do credit scoring, make hiring decisions etc. make secret use of personal data and perhaps "risk-profiling"easy to hide preventing discovery and redress.

Question 7: what legal or governance mechanisms might support ethical use of data ?

Potential mechanisms or techniques that may be of interest:

- Safe havens or virtual research environments in which data are *accessed* but not *disseminated* – see above
- Multi institution linkage and anonymisation – MILA – developed by Mark McGilchrist, Dundee – a way to separately store the identifiers and the data values, then use pseudo identifiers and federated database methods to carry out the linkage and extraction of data across a network so that only the intended, authorised recipient gets the key to unlock the linked data extracts. Ideal for linking data held by different organisations who will never agree to exchange data (eg. a supermarket and a hospital)

- Early substitution of identifiers with the same unique but irreversible random identifier per person, before each dataset is transferred to the data warehouse – to prevent insiders and researchers from ever knowing whose data they are managing. Adopted by the SAIL data warehouse in Swansea University.
- Use of agent technology to elicit, represent and negotiate an individual's preferences about what types of data they wish to share, with whom, for what types of research

Final comment: there is an assumption in the consultation document that the only real concern is about the privacy of patient / citizen data. However, in some situations (eg. when discussing research to be carried out in general practices), a significant but unspoken barrier to data sharing appears to be the sensitivity of *professionals* to others viewing - or being able to calculate - their performance. In our experience, this may sometimes be the over-riding concern, but is usually cloaked in rhetoric about patient privacy. This topic is rarely mentioned or discussed but should be, to open up this issue to debate and identify ways to overcome such barriers to useful data sharing.