Short term discussions towards longer-term information governance

1. Should the UK sequence the full genome of the entire population?

No. There are many privacy-related reasons why this is a terrible idea,[1] but the proposition is also undermined by the basic science of precision medicine. [2]

For research purposes – 100,000 Genomes, UK Biobank, etc. – there are reasons to collect samples on a prospective basis. For direct care, i.e. working out the condition that a patient has developed, a historic sample is of little use.

Just as a current blood test is required to determine current blood chemistry, an updated genetic sequence will be needed to identify relevant proteins, and which genes are active and which are not. As genetically targeted medicines become more available, this will become a much more common case – as the genetic sequence taken before a drug was administered will say nothing about what actually happened after it was administered.

Over time, genetic sequencing will drop in cost and increase in specificity to the point that it will be considered as a blood test is today.

While the results of blood tests are kept around, the biological material itself is destroyed – because otherwise there would, over time, be an excessive amount of degraded material to handle. Similarly, while technology has advanced to the point where the genetic sequence of the entire population could be stored on a lot of discs somewhere, it doesn't mean that anyone would gain any benefit from that – beyond companies that sell discs.

In practice – in the long term – as new tests become available, they should be integrated into the existing blood tests process, and the outputs should be stored. And if a patient wishes to dissent from any particular test – whether analytical or predictive – that is entirely up to them. In this way, genomics and proteomic medicine could slot in *entirely within the current information governance framework*; just as new blood tests don't need a redesign of the blood service. This approach includes the promising new blood tests for cancers, which need to be run on a contemporary blood sample.

While patients with particular conditions may choose to opt in to a particular research dataset – or choose to maintain a copy of their cancer's genome for future diagnostic purposes – there is no need, and must be no requirement, for anyone to do so.

The principles behind all three Caldicott Reports then stand without modification. Separating research and direct care was done for a reason; it means that patients get to make two informed choices, rather than one conflicting choice. Conflating the two is, and always will be, a terrible idea.

---

[1] Not otherwise covered here, but for a taste, see: https://medconfidential.org/2017/the-home-office-secretive-invasive-and-nasty/

[2] Q178 onwards http://data.parliament.uk/writtenevidence/committeeevidence.svc/evidencedocument/science-and-technology-committee/genomics-and-genomeediting/oral/48729.html

2. Can there be innovative, speculative analysis of individual-level sensitive data in a way that is Consensual, Safe, and Transparent?

Yes.

If we already knew the outcome or side-effects of a research programme, it wouldn't be innovative research. That said, individuals who wish their data used should not be subject to risk simply because they are happy for research to be conducted, safely and transparently. The environment in which innovative, speculative research and data analysis takes place must be designed to constrain any possible harm – by limiting what can be removed from the environment itself, i.e. by the use of a safe setting.[3]

Mistakes and accidents – such as breaches of small cell counts – can be mitigated by removing the ability to cause harm. (And, in cases of spear-fishing expeditions, they are entirely obvious when audit logs are examined.) David Beckham's metatarsal treatments may uniquely identify his medical records, but should you find those records in a safe setting environment, there is no way to use that information – and even looking is easily detectable by a third party.

Some innovative projects are larger than others; Genomics England runs research projects alongside direct care, and it can afford to fund the time it takes to get a patient to give informed consent. But in order to keep its own promises, it locks the data away in a safe setting, as "Inherently, patient-level data can not be treated as anonymous."[4]

Other projects with lower risk / publicity profiles may be able to fund less time, but should still adopt the same principle of using a safe setting for data analysis. If the setting is safe, the assessment of risk within the data become less relevant – not forgetting the risk assessed by the patient is often quite different to the risk assessed by institutions – because the content of the data are not designed to provide any protection; the protection comes from the safe setting, and the procedures required to access it.

Some industry lobbyists object to safe settings, as they make misuse harder. But safe settings in fact make high risk analyses safer, allowing for more higher risk analyses to be done. *Bona fide* legitimate research can continue to be conducted, following the principles of academic inquiry and curiosity-driven science. Projects run by press release, however, may be as harmed as they are harmful.

---

[3] One that meets the UK Statistics Authority requirements for such an environment.
[4] Q130 - Genomics England http://data.parliament.uk/writtenevidence/committeeevidence.svc/evidencedocument/science-and-technology-committee/genomics-and-genomeediting/oral/48729.html

3. Is there a need for "AI exceptionalism" in data handling and administrative data?

No. While 'administrative data' is collected as a side-effect of administrative action, and is therefore seen as a more accurate and more complete source of information about a service than other data sources, it should be remembered that every row – every detail – is information about a human being, that they did not choose to offer up for analysis.

The census has wide support and acceptance, in large part because citizens are in control: they fill in the form with the information they choose to give, from the options provided. Administrative data removes such control from the data subject. In many cases, providing inaccurate information on an administrative form is a criminal offence. And there is a fundamental difference between data handling for operational choices, and data for research or statistical knowledge.

As was seen with the controversy around the 2016 Australian census, when this is not entirely separate – and, critically, when these separations are not *seen* to be entirely clear – key datasets underpinning the statistical infrastructure suffer a crisis of confidence.

Those who seek 'administrative data' sometimes forget that it is also, in law, under the Data Protection Act, personal data – and often far more identifiable than other data sets – so the effects of misuse are far more damaging than other types of data. Much administrative data comes from (effectively) single provider services – the NHS, schools, JobCentres, DVLA – which is what makes it so valuable, and so potentially harmful. A citizen may be vague with a survey, or with neighbours; similar vaguery with DWP is punished as a criminal offence.

Without a regime for secondary uses that recognises the true nature of the data, individuals are forced to choose: do they want to receive support while finding a job, or do they want to do without and avoid the risks posed by ill-conceived data sharing?

Digital tools allow for new models of data analysis, new models of data collection, and stronger models of consent. In a digital world, it is possible to show each individual how their records have been used,[5] and what their choices are. If Government wishes to operate in informational terms as a "single Government Department"[6] under a Data Controller in Chief,[7] then it must offer a citizen view of Government[8] – so that no citizen has to understand all of government before using a Government service; they can simply understand how Government provides the service or services they have chosen to use.

No one can predict today what information processing will look like in 100 months' time in a hospital, a Department, or any organisation you care to choose. But whatever it looks like, it will look very similar to information processing there in 98 months' time. The best way for a member of the public to understand changes that may happen in the future is to show them

[5] https://medconfidential.org/for-patients/your-records/
[6] https://medconfidential.org/2017/on-what-principles-will-data-be-used-in-the-single-government-department/
[7] https://medconfidential.org/wp-content/uploads/2017/04/medconfidential-data-sharing-summary-april-2017.pdf
[8] http://www.infiniteideasmachine.com/2017/04/what-does-a-citizens-view-of-government-look-like/

how those changes would affect those processes they currently understand - that they have already used. It is therefore necessary, in order for people to understand how their data could be used with new technologies, for them to understand how it already is used, for the engagement they already know and understand. “For the doctor’s consultation you had last week, here’s how it will be different next time...”

With that approach, it really doesn’t matter whether the new technology is AI, genomics, proteomics, or any of the technological developments of the last 200 years or next. With the individual and their current understanding at the centre of the picture, there is simply no need for AI exceptionalism; AI is just another technology, like other technologies - and people will adapt, just as they adapted to the abacus and the printing press.

Recent history provides a clear warning: “exceptionalism” for any new form of data processing – especially one which is not well understood by all sections of the public, or which lacks a clear statutory basis – is likely to undermine the collection of *all* data.

medConfidential
sam@medConfidential.org