medConfidential brief response to the Call for Evidence on "Data regulation in the age of AI and other data-intensive technologies"

We limit our brief comments to data about people – which may be defined as personal data, or misdefined as not personal data. We note that Government tends to draw a great (false) distinction between the two, because it is convenient for power to forget that every data about school attainment is also a database about the lives of children. Should any human read this, you will remember the characteristics of those you went to school with in a way no "pseudonymisation" can ever touch. The health events of a person do not change, and are a shining fingerprint in every copy of the health records that the NHS offers to anyone who pays the fees.

1 – Access

The the simultaneous other DSIT consultation on data brokers similarly confuses access and copying – this consultation appears to expect that what the other consultation calls "permission" here can be turned into a free for all. The only difference between a broker and intermediary is whether you know that it's your data they're selling.

When DSIT argues that data copying and access are equivalent, as with the expansion of openbanking, we'd encourage any openBanking inspired data copying to data brokers to start by copying the Treasury Contingency Fund into the medConfidential bank account. We trust your treasury colleagues would object in ways you can not ignore.

3 – Reidentifiability

DSIT allowed OGDs to pretend that anonymisation by simply taking names off was effective. The analogy in AI is the system prompt for the few extremely large companies will reject such a task (others are left entirely unconsidered.

An entire chapter of Goldacre Review for DHSC is entitled "The challenge of privacy in health data" and goes into great detail on the topic, including:

> "It is common for people to disclose information about their medical problems or treatments, or have it disclosed about them. A comparatively small amount of information about one individual's medical history can often be enough to create a unique match in a large pseudonymised health dataset. For many people in the public eye, this kind of information may also be in the public domain. It is commonly reported in the media that a person in the public eye has had a particular kind of medical procedure, at a particular time, in a particular location. The precise medical procedure itself need not be particularly unique for this to be sufficient to identify that individual: the combination of the procedure, the approximate dates, their approximate location, and their approximate age is likely enough to create a unique combination, even if one only considered the week in which each event occurred. Having identified the pseudonym for one individual, any untrustworthy user with unconstrained access to the remaining data could then, once again, go on to view everything else in that individual's medical record: they would be breaking the law, but depending on the setting, they may not be detected, or prevented from doing so."...

> “Importantly, the risk of re-identification in pseudonymised data also increases, as the dataset grows to cover a larger proportion of the total population. The increase in risk from larger datasets is driven by 2 important factors. Firstly, a large dataset is inherently a much more attractive target: it is almost guaranteed to contain information about the target of interest, and therefore it is a security asset of greater value for misuse, illegal access, but also cybersecurity attack (which is itself another reason to minimise the number of locations storing large datasets of pseudonymised but re-identifiable data).
> Secondly, the maths of disclosure and uniqueness mean that individuals can be more confidently identified in a dataset that covers a larger proportion of the total population of interest. For example, if an untrustworthy user, with 3 characteristics about their re-identification target, found a unique match in an NHS dataset covering a 5% sample of the population, then (assuming no additional knowledge) they could only be 5% certain that the unique match really is the person they are looking for: because there might be 19 other matches in the remaining 95% of the population that they are unable to search across. However, using data on 100% of the population, if a unique match is found, then the untrustworthy user can be 100% certain they have found the target of interest.”

We would commend to you the entire chapter (and perhaps all the others as well).

Data becomes much more identifiable with AI which has read all of the newspapers, and can recall every incident that was written up, and match it to the dated health events in a dataset. The Goldacre Review’s emphasis on Trustworthy Research Environment is an important countermeasure, even if the choices of other health institutions directly undermine it.

Similar arguments apply to proteomics data being published openly on the web for anyone to use as they like, and can be matched to known genomes. Proteomes are now as identifiable as genomes, yet are openly published by research institutions.

4 & 5 – Ignored rights and ineffectiveness of legal frameworks

Data subjects should be able to see a complete list of how a data controller has processed, used, and disclosed person–level data about them, along with a single click button to turn off those uses which are dissentable / objectionable / voluntary.

With all the promise of data processing, this is something the clams made for more processing should make trivial, yet those who insist the public support their work but shouldn’t be actively informed about it never seem to do.

If an organisation claims unwavering public support, what are they afraid of in giving the public the tools to withdraw such support?

It is invariably the case that the most creepy and least ethical / honest actors will be the ones who use the new capabilities the most and soonest and want the least transparency over their actions.

Every supposed safeguard in legislation will be subverted and sidestepped by inappropriate actors in practice – e.g. the recent drunken walk by the Department of Health in England on "inappropriate access" to medical records and enabling creeps, voyeurs and ghouls.

medConfidential