Written evidence submitted by Dr Richard James Acton (DDA0040)

My name is Dr. Richard J. Acton (MSci, MSc, PhD), I am a Post-doctoral researcher, computational biologist & Free/Libre Software advocate. I am submitting evidence to this committee as I believe that I have expertise relevant to this area of policy, I am doing so as a member of the research science community and as a private British citizen but not on behalf of any specific organisation or institution.

I recently read the government policy paper: "Data Saves Lives: reshaping health and social care with data". Broadly speaking I applaud the measures that the government is proposing to take in this paper, there are many much-needed reforms outlined here which will be of substantial benefit to all. I do however have a few specific concerns largely related to questions of governance, software licencing and data ethics.

Software Licencing

I would question the use of the MIT and OGLv3 licences mentioned in [Section 7](https://www.gov.uk/government/publications/data-saves-lives-reshaping-health-and-social-care-with-data-draft/data-saves-lives-reshaping-health-and-social-care-with-data-draft#helping-developers-and-innovators-to-improve-health-and-care) for all software produced for or by public healthcare organisations. These permissive software licences permit the use of code released under them without requiring that code which derives from, includes or extends it also be licensed openly. This may be less suitable particularly for core infrastructural components of code produced by or for the government.

Code released under 'copy-left' licences such as the [GPL (GNU general public licence)](https://www.gnu.org/licenses/gpl-3.0.en.html) or [AGPL](https://www.gnu.org/licenses/licenses.html#AGPL) carries the obligation to release any derivative code under the same licence. When applied to core infrastructural components of a software ecosystem these licences promote interoperability and makes for a more competitive market for software development. If too much core code is permissively licensed the balance of a software ecosystem can tip towards too much proprietary code that stifles innovation by smaller contributors, limits transparency, and can lead to lock-in to proprietary solutions. This often happens when important and useful extensions or usability improvements are developed as part of proprietary extensions to permissively licensed code. The free and open-source software community has many examples of where the choice of a suitable licence has been essential to the longevity and good governance of community software projects I would give serious consideration to licencing at least some core code under copy-left licences. These licences protect the public investment in software systems by ensuring continued community ownership of these software tools and their derivatives. Copy-left licences also improve the negotiating position of public institutions when

contracting for software development by serving as a pre-commitment mechanism to the terms under which software developed for them must be licensed.

Creating infrastructure in the software commons can also be a valuable contribution to foreign aid as governments with a need for data management and sharing platforms for healthcare and beyond can benefit from the work done here. The wider the international adoption of common data infrastructure the greater the benefit to researchers who can request access to larger and more diverse datasets. Indeed a number of nations including India are already making use of and contributing development efforts to projects such as [GNU Health](https://www.gnuhealth.org/index.html) as a part of their health data infrastructure. These solutions put the integration of health and social care data a the core of their data model and emphasise enabling preventative care before the development of serious disease. The structures used facilitate the integration of health data with geographic and socioeconomic data to identify structural and systematic causes of ill health. This facilitates policy interventions outside of direct healthcare measures as well as informing practices like social prescribing. They can also make use of robust encryption standards, (OpenPGP), approved of by industry and privacy advocates alike.

I would suggest that the government draw further inspiration from the Taiwanese government's approach to public software projects in numerous regards; from their relationship with the g0v (pronounced gov-zero) 'civic hacking' movement to the Taiwanese presidential hackathon, their reverse procurement model for private work on public software, and their public cloud computing infrastructure to name just a few. Under Minister Audrey Tang's leadership, Taiwan has become a model for digital democratic innovation, from which all governments could learn a thing or two.

A clash of data ethics cultures, practices, & realities

The contrast between the terms under which data is acquired by private entities under EULAs (End User Licence agreements) and by which data is acquired in medical settings via patient consent forms is often quite stark. EULAs typically grant immense scope for companies to do essentially as they please with customer data except where bound by statute, and whilst disclaiming all possible liability, as is to be expected given their incentives. The situation is quite different, at least in theory, in the context of patient consent for use of their data for research.

Specifically, the option often presented to participants in clinical studies to withdraw consent for the use of their data at any time is often challenging to enforce in practice. For example, summary statistics computed on a set of patient data are not recomputed in the published literature if a patient withdraws their consent for the use of their data. Indeed recomputing results if a patient changed their consent status could in theory change the outcome of a study, the ability of a study to come to a reliable conclusion in the first place, and clinical recommendations arising from it. It practice retro-active withdrawal of consent for the use of patient data is typically not applied to published products of that data. A similar situation holds for other data products such as machine learning models trained on patient data these are computationally expensive to make and may need to be

subject to validation and regulatory approval for use in contexts like diagnostics which make retaining them extremely costly if a patient withdrew consent for the use of their data in the training of such a model.

What should be the scope for retroactive withdrawal of consent to use data? Current wording in the medical context tends to imply an absolute right to withdraw consent. This does not reflect the reality that once published in any form data may never go away if third parties have copied it, nor does it handle the implications for the withdrawal of consent for data products.

In private settings one typically relinquishes all rights to one's data with little to no recourse if you change your mind, this is also not a desirable state of affairs.

It would be reasonable to permit people to withdraw the consent for the use of their data in any new projects subsequent to the withdrawal of their consent, though this is still not without problems for studies reproducing and/or evaluating the analyses performed in previous studies for which the complete original data is needed.

Clear communication of the practical limits of the retroactive withdrawal of consent for data use to the subjects of that data. Clear systems for how withdrawal of consent is to be communicated to parties with access to the data so that it can be respected. Data already released to researchers often does not carry with it any strong guarantees of a chain of communication that would permit clinicians and trial administrators to contact those researchers with copies of the data to stop using that data. Public consultation on these points is needed to inform policy about sensible defaults, public expectations and the stringency of standards for data stewardship necessary for different data types. There are also potential conflicts between data retention policies which may mandate data be retained for a certain period and participants requests for data deletion - which should take precedence under which circumstances? In neither public nor private contexts is adherence to best practices adequately policed or enforced, in substantial part because there is often a lack of clearly defined best practices that can be practically adhered to, and a lack of mechanisms of enforcement. I am not a legal expert this may need new statutory measures or just enforcement with much more resources and bigger teeth.

Projects which facilitate and encourage keeping data in a controlled environment such as OpenSAFELY may be a part of the solution to some of these problems if coupled with enforcement actions for breaches of data ethics standards. Importantly for such environments to be practically usable by researchers they must have access to heterogeneous compute and data storage resources which is challenging to achieve securely.

As the government's behavioural insights team 'The Nudge Unit' would likely attest it is important for anyone working with sensitive data to have good defaults and systems which make it as low friction as possible to adhere to best practices in its stewardship in order to minimise data incidents.

To this end platforms like OpenSAFELY should look to other tools for open collaborative and reproducible data science work such as the Swiss bioinformatics institute's Renku platform to inform design of their data analytics environments.

For more sensitive data the use of approaches based on homomorphic encryption which permit the analysis of encrypted data without direct access to the data by the analyst may be an option, as data ex-filtration is always possible in an environment where someone can see the raw data.

Maintenance of open-stack public cloud services for development, testing and public interest software projects (as Taiwan does) and requirements around interoperability between this platform and any private cloud infrastructure would be in the government's and the public's best interest, with respect to ensuring that the taxpayer is not locked into any specific private cloud provider and can practically exercise the ability to switch providers at minimal cost. The ability to be 'cloud agnostic' will be an important component of computing procurement policy with potentially substantial financial implications going forward.

The importance of the point that you cannot leak data you do not have cannot be understated, to this end we should strongly discourage the possession of unnecessary personal data.

I would suggest that we should aim for a legislative environment where possession of extensive databases of personal and personally-identifying data which is not well secured and without the means for people to see what data an organisation has about them, any products of that data, and to withdraw consent for its continued use including requesting its verifiable deletion would be considered a toxic asset.

3. Decisions for the data commons

The current proposals in the policy paper for how decisions should be made about data sharing agreements seem to me to be lacking in appropriate accountability mechanisms. Datasets such as aggregate health data from the NHS are a publicly held data commons and should be managed in accordance with the principles for successful governance of a commons laid out by Elinor Ostrom in her Nobel prize-winning work on the subject.

There should ideally be direct public representation in data trusts with the power to block any data-sharing agreements to ensure that any such arrangements are in the interests of the individuals in the data trust and/or reflect the preferences and intent of the members of the data trust for the uses to which their data is to be put.

They should also ensure that such arrangements are clearly and transparently communicated to the public before and after they are agreed.

When making decisions to publish data careful consideration to the risk posed by de-anonymisation attacks, which can reveal the identity of individuals whose data was published in a form that was thought to be anonymised or pseudonymised, should be given.

In summary, I would appeal to the committee to consider:

- Licencing core software under a 'copy-left' licence like the GPL
- Consulting with the public about good data sharing defaults for specific data types and establishing their expectation with respect to privacy and withdrawal of consent to use their data.
- Improving accountability and enforcement for best practices in data stewardship
- Minimising the amount of unnecessary data held
- Design systems for persons who have access to and work with data so that it is simple and practical for them to adhere to the best practices for data protection.
- Having direct public representation in important decisions about data sharing with the power to block agreements.

January 2022