medConfidential comments on the Wellcome call on The National Data Library (for personal data)

There are many possible technical architectures for the National Data Library – the NHS alone currently has five possible base architectures (in alphabetical order: FDP/central care record, NHSE's national SDE, OpenSAFELY, regional SDEs, TREEHOOSE/etc from HDR)[1] and ONS has the IDS (pg 123).[2] Irrespective of whether any submissions are a "five safes" model or any other, the tenets of consensual, safe, and transparent should apply to them all.

Every project should demonstrate how it complies with the transparency standards for publishing a comprehensive list of all projects in the system. This can range from a real-time OpenSAFELY jobserver to Parliamentary Questions about Palantir or the Integrated Data Service.

Transparent – Public confidence

The NDL will not be health first, but it is likely that Government will want to use health data within it. How the NHS's National Data Opt Out is given effect is a policy question (NDOO is full of loopholes and NHSE will generally find one if they benefit from the project).However, it's inevitable that some projects will need to respect the NDOO, while others won't. The system must address this issue in a consensual, safe, and transparent manner.

For the NDL data from the rest of the government, while the gov.uk Personalisation push continues to generate more hype than practical implementation, this personalisation could be leveraged to build trust in demonstrating how data is responsibly utilized. It could also showcase how your data has been responsibly used and highlight the benefits that have resulted from it.

HMG will at some point use the Gov.UK Login to create a "rest of government opt out" – if not initially, then after the endless scandals. When "tell us once" becomes the banner for gov.uk integration, is it feasible for government to require anything else for dissent from unnecessary processing or objecting to being used as a human subject in data driven research?

ONS used to seek to maintain public confidence, and that involved data minimisation – providing only the data necessary for the analysis and enough data for statistical validity while not increasing risk for population scale coverage. That should continue.

---

[1] All of these predate the central care record proposal which removes the "unique" offer to research of the "regional SDE" approach. Central care records will also contain raw GP medical notes, which "research" (in this case specifically, AMRC/Biobank/HDR currently argues it does not need and does not want However, it is unclear whether that position is tenable or credible where notes are available – ? Is the "no" position tenable when they could have them and a researcher comes with a project? What are the projects after those?. Would biobank really not want them and really not decide to hand them to a bunch of eugenicists?

[2] medConfidential's remarks at the Scrambling for Safety event on 28th November go in here. But this PDF was published a week before the event on Euston Road so no link as yet. You should come.

Consensual – Data inputs

Admin data is not voluntary, will the NDL care? Will anybody who doesn't want their data to use for purposes beyond the initial service be forced to be included in a library they don't wish to be in?

Unless structural, technical, and legislative measures are taken, it trends towards inevitable that all data will be in the NDL in the long term; political promises otherwise are temporary, therefore the NDL will have your DNA in it as another eugenicists paradise.

There is a temptation for ONS, especially recently, to treat admin data as equivalent to a survey. It is not. ONS argues that the public and business trust ONS as they respond to surveys (a lot less than they used to), but even ONS treats the census as different due to the statutory compulsion and prosecution for not filling it in. The same compulsion applies to what becomes HMRC's PAYE/income data, or DfE's school returns.

Health data (and other admin data) and ONS survey data are not all the same – there are large structural differences.[3] If you don't fill in the DWP forms for UC, then you can't feed your kids, or yourself, or keep a roof over your head. In what way is that voluntary submission like a survey? If you don't like a question on an ONS survey, even the census, you can give a convenient/funny answer ("Jedi"); if you do that to DWP they threaten you with jail.

Justice data will be included in the NDL, which while some of it is about criminals and perpetrators, it is also by necessity about victims and often the worst thing that happened in their lives. Inclusion of data without any choice is another case of re-victimising victims, and it is not unknown for the state to be complicit in the original acts.

Safe – Good TREs Do Work

It only takes one unsafe process to undermine all others

Good TREs work and can be a benefit to research; a bad environment is toxic. When designing standards, we would never have envisaged that an environment would allow commercial users to keep the code they run secret from the environment owners, but HDR decided that was an important feature, just as they decided they should allow arbitrary users to run arbitrary simple analyses on sensitive datasets because it met their narrow immediate goals.

Some projects may seek to undermine the "safe' principles by returning to special pleading that they misname "safe return" – in which a researcher daydreams of research impact so important that it does not need peer review, does not need external scrutiny, does not need publication, does not need writing up, but they must be able to immediately contact remote unknown clinicians who do see patients and instruct them to change the care they deliver. It

[3] We wrote this in 2021 to help NHS England understand their ONS friends; it was written for an NHS data audience: https://medconfidential.org/wp-content/uploads/2021/12/ONS_NHS-analogies.pdf

is a HDR fiction entirely divorced from reality. Where a research finding is clinically relevant, the research should be written up (even as a preprint), and the publication with steps for replication should be shared with the data owners who can independently check the analysis using their own data. There is no need to undermine the five safes approach because some ageing researcher dreams of winning a Nobel someday and doesn't want to listen to others. Anyone with clinical findings can publish, with far less false positives or risks of harm. Any organisation that suggests "safe return" as a credible approach doesn't understand the problem.

It only takes one unsafe process to undermine all others

Whether a TRE is "good", and whether the NDL will be "good", is all in the details.

We look forward to reading and assessing the submissions published.

medConfidential
November 2024