## medConfidential more detailed submission to the First Goldacre Review

The white paper[1] promotes and repeats all the enablers of the mistakes of care.data, and the Goldacre Review may be cited in support of those mistakes, even if that is not the author's original intent. NHSX was not around to learn the lessons that NHSDigital learnt, and NHS England chose to ignore, phrased in one academic paper, as "You hoped we would sleep walk into accepting the collection of our data",[2] or as a Nature paper put it "You should at least ask".[3]

### Data flows in and around the DHSC family

The status quo of NHS data for secondary uses is full of entrenched interests and disingenuous lobbying, including by national bodies.

The ToR for this Review are correct that the flows around the family are complex, and could be simplified. One obvious existing statutory approach would be to merge all secondary uses to flow through the statutory safe haven – this is likely to be optimal for GP data. We note the argument that specific data uses may have divergent benefits or diseconomies of scale. This Review can not resolve that debate, but what it can do is provide a better evidence for when the debate continues:

> **Recommendation: A DHSC family data flow diagram.** One of the earliest tasks that the newly formed MoJ Digital team did was map the entirety of the data landscape[4] of the flows of people through the criminal justice system, and how they interacted with many different services. Without analogising researchers or NHSX to criminals, **DHSC should commision an equivalent graphic of how all the different places in the widest extent of the DHSC family who may make data available for research (and other secondary uses), and the flow of requests and data between them**. The diagram should also include a reference of how many and have a modern data infrastructure, link to the DPIA, produce accurate data release registers, and publish details of successful applications.

---

[1] https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/960548/integration-and-innovation-working-together-to-improve-health-and-social-care-for-all-web-version.pdf

[2] https://pubmed.ncbi.nlm.nih.gov/26280642/ Sterckx S, Rakic V, Cockbain J, Borry P. "You hoped we would sleep walk into accepting the collection of our data": controversies surrounding the UK care.data scheme and their wider relevance for biomedical research. Med Health Care Philos. 2016 Jun;19(2):177-90. doi: 10.1007/s11019-015-9661-6. PMID: 26280642; PMCID: PMC4880636.

[3] https://www.nature.com/articles/ejhg201630 McCormack, P., Kole, A., Gainotti, S. et al. 'You should at least ask'. The expectations, hopes and fears of rare disease patients on large-scale data and biomaterial sharing for genomics research. Eur J Hum Genet 24, 1403–1408 (2016). https://doi.org/10.1038/ejhg.2016.30

[4] https://mojdigital.blog.gov.uk/wp-content/uploads/sites/58/2015/11/criminal-justice-services-landscape-map.jpg via https://mojdigital.blog.gov.uk/2015/12/01/opening-up-data-in-the-criminal-justice-system/

**The National TREs**

In Summer 2021,[5] medConfidential will assess the various 'national' TREs in use across health data (and wider public sector). We'll include the NHSD's TRE, ONS, SHIP, and SAIL, GeL and any other temporary TRE that makes claims of wider access (the HDR attempt will likely be included), and the confidentiality, safety, and transparency measures taken by each, and how they compare. This is expected to also include the extent to which synthetic data is available for use, data is sampled or a census, and the technical measures stated as in use, and actually in use, and the consent status of data within the system to the extent that it interacts with other measures and the supply chains (and shareholdings) of those who have received medical research money for TRE work.[6]

> **Recommendation: TRE managers must share with NHS Data Controllers full information of their supply chain (and governance thereof).**

The vast majority of analyses are common and routine – this month's statistics, compared with last month's statistics, and compared with next door's statistics. All of these can be automatically produced, either in bulk as ONS does with the census, or through easy to use interfaces at NHSD such as HDIS. The statistics, once produced, can be compared in any number of different visualisation tools and packages.

The data curation programme is a money pit which will deliver nothing meaningful. It is a revenue stream for expensive consultants without any material effect on care or research. The DeepMind/Moorfields project shows how to make this work better – spend time with clinicians to co-create research and innovations. Similarly, the approach taken, and apparently mandated by HMG, for digital innovation hubs to have mandatory commercial partners is a recipe for failure and data harms. It is unclear what the programme has achieved, and reading the Farr/HDR annual reports back to back, the model of "commercial entities can do it" is something which has largely perpetuated the status quo at significant ongoing expense to thee public purse, rather than anything which may have reduced the overall cost to the public.

The Government's proposed 'PFI for data projects' will fail – it will cost as much money as Government can be persuaded to pour down that particular drain, and the data quality will not change after the pouring stops. That is not to say that entities will not profit from the pouring, they almost certainly will, it just won't do anything for sustainable research.

But the headline is that **any TRE is better than dissemination, because *anything* is better than dissemination.** As the UK pursues trade deals with other countries, which may include dissemination without regard to national boundaries, the only available safety measure in that world is for there to be *no* dissemination at all – TREs only.

---

[5] if the COVID19 vaccine works

[6] E.g. https://find-and-update.company-information.service.gov.uk/company/SC324508/officers

**Reidentification Risks**

For new stories over the last decade, we have a long month-by-month timeline published on our website,[7] with an additional detailed timeline on the Deepmind/RFH misstep,[8] and a selection of some of the most egregious cases.[9]

On news reports that illustrate the risk of reidentification:

- Tony Blair's heart trouble - dates, hospitals, procedures
- Michael Gove's son - date, injury, hospital
- David Davis' nose - broken five times
- Nick Clegg's partner's broken elbow - date, hospital, injury location
- Kate Winslet's 3 children's birth dates - one emergency cesarean, hospitals

Plus Latanya Sweeney's classic reidentification of Massachusetts' Governor Willam Weld[10] and her 2019 Nature paper.[11]

With so many mergers / takeovers in the health data space, the risk of 'contract swaps' providing tech corps with health ambitions access to bulk NHS data[12] (N.B. If you're bored of Google, try IQVIA which - as IMS Health - got approved to link HES and CPRD data, which it's merrily offering to US customers.)

When discussing the 'reassurance' aspects of the Review, it is important to note that past 'reassurance' would need to exist: Single strike never enacted[13]; NHSD TRE appears to have been minimally used.

A TRE can encourage or mandate transparency due to the need to check outputs. The lack of transparency as to how "QCovid" works seems unwise, given the bad interaction that "QRisk" previously had when it crossed from one GPSoC system to another;[14] that mistake has been repeated in covid.[15] Such mistakes are not unique to health, and are only resolved by the ability of others to scrutinise analyses and processes.[16] [17]

---

[7] https://medconfidential.org/information/media-coverage/
[8] https://medconfidential.org/whats-the-story/health-data-ai-and-google-deepmind/
[9] https://medconfidential.org/for-patients/major-health-data-breaches-and-scandals/
[10] https://arstechnica.com/tech-policy/2009/09/your-secrets-live-online-in-databases-of-ruin/
[11] https://www.nature.com/articles/s41467-019-10933-3
[12] https://techcrunch.com/2019/10/22/google-has-used-contract-swaps-to-get-bulk-access-terms-to-nhs-patient-data/
[13] The Statutory Instrument placing the Confidentiality Advisory Group on a statutory footing to advise NHSD has never been laid.
[14] https://www.dailymail.co.uk/health/article-3585149/Up-300-000-heart-patients-given-wrong-drugs-advice-major-NHS-blunder.html
[15] https://www.liverpoolecho.co.uk/news/liverpool-news/invited-covid-vaccine-because-nhs-19857990
[16] https://theconversation.com/the-reinhart-rogoff-error-or-how-not-to-excel-at-economics-13646
[17] https://theconversation.com/excel-errors-the-uk-government-has-an-embarrassingly-long-history-of-spreadsheet-horror-stories-147606

## What do to about data risks - repeat the cancer registry migration for other centres of expertise in particular data silos

Following the model of the cancer registry migration to NHSD, the statutory safe haven alone should be the process for access to data. Replicating the approach of the cancer registry move, other areas of speciality (such as CPRD) should support and scrutinise applications for data within their remit, as part of the DARS process.

CPRD today publishes details of some projects from their customers, which is the outcome of those who, for example, say:

> *"The aims of this study are therefore to **identify and describe ... patients** and to assess their burden of illness."*
> *"this study aims to **identify patients ... using CPRD and a patient questionnaire**"*
> https://cprd.com/protocol/assessing-burden-illness-generalised-refractory-myasthenia-gravis-england-using-cprd

or:

> "**we will identify patients** who were diagnosed with von Willebrand disease. We **will send GP questionnaires for a sample of patients with von Willebrand disease to confirm the diagnosis and obtain information on treatments that are not captured in the data**,"
> https://cprd.com/protocol/epidemiology-von-willebrand-disease-uk-1989-2016-prevalence-treatment-patterns-descriptive

While CPRD may claim that their dissemination without a TRE is safe, any claim that re-identification is impossible is demonstrably false by the actions of other CPRD *approved* projects. If data can be reidentified to allow a 'patient questionnaire', it can be reidentified by other means. CPRD's insistence on sending data overseas means that those who re-identify outside the country are not within the UK legislative protections against re-identification (such as they are).

We therefore note, with no surprise, that the NHSD audit of CPRD showed CPRD does not have a DPIA.

It is unclear how CPRD considers it is acting lawfully in disseminating GP data overseas for reidentification without doing legally required steps, but given competence of government advice on data,[18] perhaps MHRA/CPRD has followed the Secretary of State's lead in believing the law does not apply to them.[19]

---

[18] https://twitter.com/medConfidential/status/1357037423172141061
[19] https://www.wired.co.uk/article/nhs-test-and-trace-unlawful-data

**What works**

For routine outputs used in analyses, the approach ONS takes to the production of massive numbers of standardised census tables, on a basis that is almost entirely automated, should be replicated for large parts of the NHS. That includes the micromanaging of the hospital as NHSE may desire.[20]

Despite the desire of secondary users and medical researchers, the purpose of a clinician is to care for their patients, not to curate better data for the research of others.

However, the DeepMind/Moorfields project made it clear – when clinicians understand what data quality is needed for cutting edge AI uses, they will meet those standards. When it isn't needed for that, they will meet whatever the normal needs are. The DeepMind/Moorfields project needed around 2 weeks of images for their work; that is neither a lot of data, nor hard to recreate as needed when there is a meaningful partnership between the NHS and additional researchers.

Expecting random other entities to be able to do cutting edge abstract analysis is unrealistic – what works is partnerships.

Is there any evidence that the LHCR mandate requiring patient data to be given to commercial third parties for 'analytics' has provided any meaningful benefit for the NHS? DHSC should be required to provide some in response to the Review. Similarly, the model of HDR of mandatory commercial entities for 'commercialisation' has not seen meaningful benefit to the NHS, and potentially is a political risk for ministers who seem to turn a blind eye to the cronyism in data choices.

Addressing data supply chains, there are examples[21] where data access for one uses has been repurposed for profit by others. Every analysis that is used in decision making or procurement should be accompanied by an analysis and input certificate showing when the analysis was done and the pedigree of the data used.[22] Those certificates should be produced automatically by a TRE, and included with analysis outputs to show that proper IG had been followed at every step.

Matt Hancock got it right with vaccines[23] - proving that when there was a priority besides cronyism, he would make the right decision. Will the same happen for data?

---

[20] Noting the original cause of the mid-staffs scandal was a small number of metrics which then got gamed by consultants who could be paid to explain how to game the metrics. See the footnotes in the letter https://medconfidential.org/wp-content/uploads/2021/02/2021-02-CDEI.pdf

[21] https://www.whatdotheyknow.com/request/ai_agreements_with_orthai

[22] https://medconfidential.org/2020/analysis-and-inputs-reporting/

[23] https://www.spectator.co.uk/article/secrets-of-the-vaccine-taskforces-success

## Palantir

If what the NHS uses Palantir for are things *only Palantir can do*, then there may be a case for its continued use; the public debate around this is characterised by PR babble, technical bluster and political posturing on all sides.

The Review should recommend the Boards of NHS Digital and England jointly commission an evidence-based independent assessment of: a) what the Palantir/Data Store was used for; b) what it *alone* could do during the pandemic; c) what NHSE expects it will do after stepping down from the Level 4 alert; and d) whether Palantir is the most cost effective way to do those things. This assessment should be led by a person with experience of being a Caldicott Guardian of a large or national NHS body that is not NHS England.

NHS England asked Palantir to build the Data Store as it needed a solution rapidly, while officials were fully aware that Palantir's reputation is toxic to public confidence. Unlike some other fields, however, the NHS has a single urgent and pressing need: it has to save lives. It is that single urgent and pressing need which made the NHS response to COVID-19 so effective. This does not exist in the same way in other policy areas, where there are often complex stakeholders with contradictory needs.

As we move into the post-pandemic world, the NHS/DHSC should ask those with interest in the Palantir decision during the pandemic a simple question: what *should* the NHS have done in that circumstance? (given that letting people die was not an option).

**Other points  around the whitepaper**

We'll cover these points in our whitepaper response, which will be out by the time your report is, but which isn't yet available:

- Critical importance of independent decision making on secondary uses – NHSE may think it did a good job, only because it hasn't told anyone what's going on who might point out their mistakes. Did they hand patient data to a sham organisation because they were under political pressures to be quick rather than careful? Without transparency no one knows, and NHSX/NHSE/DHSC will be eternally terrified that they've made the same mistake as doomed care.data… Data controller decisions should only be taken by the statutory safe haven, which should continue to be independent and make decisions which are avoid beyond the current spending settlement horizon of NHSE politics

- **Patient visible logs of which organisations have accessed and when.** We draw your attention to the 'Creepy Single Doctors' problem[24] which is adjacent to the reidentification debate. You may also want to consider a Centralised dissent function (which will make the inevitable merger of the two SCRs as one easier) – there can not be public confidence in research if patients have no meaningful way to dissent from data uses. If NHSX idolise the Facebook model of making it too difficult to dissent, that undermines confidence in research. It is a tenet of research ethics that

[24] https://medconfidential.org/wp-content/uploads/2020/09/Creepy-single-doctors-v2.pdf

you don’t experiment on any individual who doesn’t wish to be involved, and that applies to data as much as to any clinical intervention.

We’re happy to work with you further as you may find helpful

medConfidential
coordinator@medConfidential.org