medConfidential comments on UPD's "What happens to health data"

Previous 'incarnations' of Understanding Patient Data [1] would most often kindly have shared a pre-publication copy of a document like this with us privately, and medConfidential would have given comments on that same basis – leading to a better product for patients and the public, and helping to improve patients' understanding of data.

The current incarnation of UPD didn't do so.[2] So we are publishing what we would have said.

Page 3: *"Healthcare research and planning often use data gathered from limited numbers of patients"* – it is unclear to what activities this is referring; some research studies can indeed be small, but planning is often done using population-scale datasets.

Page 4: Confusingly, this starts out by talking about "notes", but later refers to what would appear to be values for the various readings taken. Which are not "notes".

Page 6: Again, vague use of "notes". A person's medical history may contain clinicians' free text notes, coded data, lab results, readings, even attachments (e.g. scans and letters). When explaining health data, and the different types that exist and are used, it helps to be clear and accurate.

Page 8: What on earth is *"indirect* use of data"?! Someone looking at a screen obliquely? Data collected for one purpose may be re-used for another – this is called secondary use. The data is either used, or it isn't. There's nothing "indirect" about it, unless this refers to patients **not being told** how their data is used?

Page 10: Use of a fictional example is unhelpful, not least because it allows one to state whatever one wants as if it were 'true'. In reality, the majority of large datasets that are used include much more data than the stated 'example', simply because *anything* in them might be useful. If researchers knew that only that precise list of variables would be needed for their research, it wouldn't really be research.

Page 11: This page by itself implies that the biggest harm from a large health dataset is that some people don't want to be in it!

Page 12: This page fails to mention the single biggest harm of large data sets, which is that they – especially if poorly managed and poorly communicated – can undermine public confidence and patients' trust. And if patients cannot trust that what they tell their doctor will be kept in confidence, they may not reveal things that could significantly impact their own or the public health. To talk about paying patients for use of their data, but not to mention trust is a gross omission.

Page 13: How does this made-up example align with the one on page 10? What useful information does it even add, e.g. why no mention of how pseudonyms facilitate data linkage

---

[1] We understand why Wellcome largely defunded UPD; since NHSE & DHSC clearly weren't going to take on board what they said, funding UPD was not easily justifiable given the other draws on the Trust.

[2] We're not really sure who is running UPD these days, after the good people we usually dealt with saw the direction of travel and moved on...

over time and across data sets? Given this example is all made up, could it not even provide an outline of *how* a treatment might be improved?

Page 14: Some of the statements on this page are simply untrue. While they may talk a good game, HDR UK and NIHR don't run their own Five Safes TREs; if they have "adopted" them, then so has NHS Digital – which has built one, which is being used. Indeed, HDR UK paid NHS Digital to make use of NHSD's TRE during COVID – so if HDR UK gets mentioned, so should NHSD. Also, the 'Further Information' provides zero evidence of these assertions about HDR UK & NIHR.

Quite apart from the implication that *"anyone who manages health data within any organisation"* may be able to satisfy the high bar of the genuine Five Safes, "Safe data" in the disordered, poorly defined list on this page is not even aligned with the examples given elsewhere in the document. Those state that the data is pseudonymised, and only pseudonymised, which means it remains personal data. "Safe data" is about being as safe as possible, because *absolutely safe data* may harm the research questions; the only absolute safety *for an individual* is not to be in the dataset.

Page 15: The 'definition' of "anonymisation" on this page repeats many of the same weasel words and slippery constructions we've seen over the years from those wishing to justify data re-use without consent or transparency. To make things worse, the document wrongly conflates anonymisation with de-identification, and is *legally incorrect*; de-identifying data by removing the obvious identifiers doesn't make data any less personal data under the law. It is also unclear how the examples given earlier in the document relate to what is said on this page. Is UPD claiming that personal data can somehow be deemed "anonymised", and thus outside the law? Such arguments have been tried before. And failed.

(N.B. There is no debate on when information is anonymous – if data is anonymous it can be published, e.g. as statistics. If it cannot be published, it is not anonymous.)

Page 16: The definition should say that the pseudonym replaces specific variables, while leaving others untouched. (This would also make page 17 much clearer and simpler.) While this does mention that pseudonyms "distinguish between individuals", it again omits to point out how this is used to **link** individual's data together. Might this perhaps be because this ability to individuate patients within pseudonymised datasets makes it quite obvious that their data is *identifiable*?

Page 18: While there may be "a spectrum"of de-identification techniques, that does not mean that the identifiability of data itself is continuous! The point at which data becomes *genuinely* anonymous, i.e. not personal data at all, is a hard red line. Statistics are not 'on a spectrum' and, for that reason, can be published. Other treatments of personal data may be sensible, and may make the data safer to handle in some circumstances – but rich, linked, individual-level health data, useful for research and planning, will *always* be personal data.

Page 19: With a bit of work, this could be a good clear page. There are some obvious omissions, such as any mention of Hospitals or NHS Trusts – and it was clearly an oversight not to update this page since CCGs were abolished on 1st July...

Page 20: Statements on this page are a bit of a hostage to fortune, given Ministerial promises that marketers (not at all true) and insurers (probably true) cannot get access to patients' data. Is

NHS Digital content to allow UPD to make statements such as this on its behalf? Are Ministers? *"Anyone who fails to meet these criteria **will** not be given access to NHS health data"* might more accurately be stated as *"**should** not be given access"*, since that decision depends on other criteria – and, in the commercial space, largely politics.

Page 21: While SDEs/TREs are a good thing, this page seems to suggest they are the only way data can be used. This is untrue. For the latest month for which figures are available,[3] of 1,413 projects across the 461 organisations in the 2022 data release registers, 0 (zero, zilch, nada) organisations had TRE-only access. 130 organisations used NHSD's TRE once; 347 didn't use it at all. And for most of those latter projects, the "technical and legal controls" amount to 'trust us not to cheat' – while the audits that have been done show organisations regularly do break the rules, and yet are not punished.[4]

(N.B. This page once again conflates "NHS England" with the NHS in England, a rookie mistake that previous incarnations of UPD would never have made.).

Page 22: This page obfuscates that controls still largely don't exist, or aren't used, and – while acknowledging that current protections are ineffective – neglects to mention that the National Data Opt-Out was only respected around 46% of the time that data was disseminated.[5] There's also no mention that "transparency" for some datasets is limited to a pitiful dribble of information released via PQ,[6] over two years after it was first asked for.

Page 23: *"Not use data for marketing"* is demonstrably untrue – NHS Digital's data release register shows that pharmaceutical marketing is a regularly permitted purpose. And on *"Make it clear how and why data is being used"*, see notes on "transparency" from page 22. UPD may wish to link to TheySoldItAnyway.com, so readers of the document can more readily see how data about them *has* been used, rather than being pointed at a Power BI dashboard or dense spreadsheets.

The statement, *"there are ways to opt out of having most types of your health data used for most purposes beyond your individual care"* is misleading. "Most types" is questionable, and "most purposes" is by definition not true, given NHS Digital's release register shows patients' NDOOs are applied less than 50% of the time.

Page 24: "Those who have knowledge, don't predict. Those who predict, don't have knowledge." The application of *Deep* Learning to imaging may be showing some potential; the majority of Machine Learning and other "AI" projects initiated to tackle COVID delivered nothing. If the future of large datasets in and around the NHS is as credulous as this page, then we are in real trouble...

Page 27: What no NHS England, or DHSC? And is HDR UK really *"a national institute that aims to pull together the UK's health data to enable discoveries that improve people's lives"*? It's quite clear who UPD has been listening to, but is it now publicly stating that HDR acts as a data controller?

---

[3] https://goodtreswork.com/

[4] https://digital.nhs.uk/services/data-access-request-service-dars/data-sharing-audits/2021/data-sharing-remote-audit-icl

[5] https://theysolditanyway.com/

[6] https://questions-statements.parliament.uk/written-questions/detail/2022-07-08/HL1602/

Also, medConfidential.org appears to be missing.

Comment on reference to “free-text”

Parts of the document read as if Understanding Patient Data is *expecting* a push for free text from patients’ medical histories in, e.g. a future incarnation of care.data.

UPD was loosely inspired by the information and advocacy organisation, Understanding Animal Research (UAR). Cute mouse logo aside, the effect on patients of some of the things for which UPD advocates is equivalent to the effect on animals of UAR’s advocacy.

The 3Rs are good, but what’s being done is still not pleasant.

Taking the analogy a logical step further, **what are the 3Rs for health data?**

In many ways, researchers who break the rules covering mice face greater punishment than those who break the rules covering people's lives in their data. That’s not to say we should be any less strict around harm to mice, but perhaps we should respect people’s choices better…

medConfidential
October 2022