

## **What is biobank doing? Some Superficial outcomes of medConfidential's *truncated* initial look at biobank**

In a meeting with Biobank's Primary Investigator,<sup>1</sup> I offered to look at the material that biobank provided to the GP Profession and to their cohort. It has not yet been provided, although briefings have been made to the press which [we've dealt with elsewhere](#).<sup>2</sup>

Beyond GP, biobank has access to data [from the NHS](#) from hospitals (A&E, inpatient, outpatient, critical care), data on all cancer patients, mental health data, and diabetes data. Many of the issues with biobank overlap with those around [Our Future Health](#) and FDP.

In the absence of requested material, we superficially and briefly took a look at the material on the website. It is that public material which forms the basis for this interim summary of what we found. We would have shared what we found to biobank only, but given the lack of candour from biobank to others recently, this assessment is public.

### **Biobank make things up?**

Biobank and NHS England sent a letter in late September telling GPs that BMA supported an action they did not support. I do not know *how* this went wrong, but it did.

Biobank claimed support that did not exist.

This undermines biobank's point that biobank, and biobank alone, should be trusted to make trustworthy statements about what data they collect, and what data they share.

That process failure on the part of biobank (and E? DH?) suggests that when biobank say process exists to be followed, that statement may have some significant caveats.

#### **1. Is there a shared position between E+DH+biobank on what went wrong and who was responsible?**

Without a clear line of responsibility, and without clear delineation of liability, if GPs were to rely on biobank assurances like claims in the letter, GPs would be liable for biobank's mistakes.

#### **2. Is DH/E willing to take on legal responsibility for all mistakes by biobank, or do they remain on each individual GP?**

### **GP data is already in biobank**

[Biobank's documentation](#) says they have GP data on 230,000 people, saying: (emphasis added)

---

<sup>1</sup> That meeting and other conversations also gave some suggestions of what Biobank could do to move forward on a range of their "stuck" issues, including a roadmap to a communications channel via the NHS app, opening the medium term possibility of wider biobank activity

<sup>2</sup> The paragraph quoted at the bottom of page one in our fiscing should be noted by NHS England for other reasons which will emerge in future.

**“UK Biobank has been liaising with various data suppliers and other intermediaries** (including the main primary care computer system suppliers in England) to obtain primary care data for UK Biobank participants, all of whom have provided written consent for linkage to their health-related records. To date, **coded data have been obtained for approximately 45% of the UK Biobank cohort (~230,000 participants) and are now available** as part of this interim release. Details of the data providers and the coding schema used are summarised in Table 1.

**UK Biobank is currently in the process of securing access to data for the remaining cohort, mainly for participants registered with EMIS practices across England.”**

The last comment implies that it is TPP who is the one of the “other intermediaries” who acted as a data controller in determining data should be released to biobank. If the data controller had agreed, the current complaints about GPs not agreeing would clearly not be happening.

Table 2 of that [document](#) says biobank have 87m clinical events supplied by TPP and Vision.

**3. What data does biobank have, on what legal basis, what have they done with it, and to whom have they given access to GP data?**

We understand biobank has previously been told by JGPITC that what they were doing should cease.

**4. What did JGPITC tell biobank? What did both parties do afterwards?**

**Is there any assurance around which patients have consented to be in biobank, and what they agreed to?**

An individual GP can not reasonably assure themselves that the materials they are provided with are representative and accurate; not for many studies, not for many programmes, as both change over time and 8000 GPs should not each have to check everything themselves. This assurance is an activity that, in practice, most GPs look to professional bodies to do on their behalf, namely BMA and RCGP, via the Joint GP IT Committee.

We do not know exactly what patients agreed to when. It is not known, and possibly not knowable, how a GP would know that one of their patients is fully and correctly consented into biobank, satisfying their ethical<sup>3</sup> obligations, other than biobank saying so...

**5. What is the chain of trust between a patient being consented and data on them only flowing from the GP/hospital/etc.**

---

<sup>3</sup> Legal obligations could be set aside in various ways; moral and ethical responsibilities can not.

**6. How do GP IT providers get an authoritative and reliable list of who is in biobank, such that the legal burden shifts away from each individual GP and on to that process? (likely onto NHS E?)**

There is no clarity on the exact data flow between biobank, GP, and back to biobank, with all the steps along the way, and how changes are accommodated over time.

Biobank appears to have a list of NHS numbers from patients, and GPs have been given a button that says “approve biobank”. There are a lot of steps between those two.

- 7. Does a GP see which patients are involved in biobank?**
- a. Since BMA/JGPITC is bypassed, who are GPs who have pressed the button for biobank relying to have done the right thing?**
  - b. Who checks that the process is as it should be?**

**8. Who will do this for all the other cohort studies as this model expands?**

We suspect this should become a function of DigiTrials (with professional review akin to AGD/PAG).

**9. What should this process look like in a world with NHS digiTrials, for all cohort studies? (Are HDR arguing biobank *are* unique, or that they *aren't*?)**

An individual GP can not reasonably assure that materials provided to them are those provided to patients, and that biobank statements are reliable.

In the case of Our Future Health, the OFH process evolves as they remove parts that dissuade people from signing up, which means no GP could ever reasonably see what each of their patients were provided with.

**10. Is there an existing process by which a patient can (re)discover they are in the biobank? Can this apply to any other “consented” data flow?**

**11. How would biobank like to start to rebuild the perception of trustworthiness that they have undermined with their actions?**

### **Biobank’s Application process**

The complete application process for NHS England’s DARS is often described as long and arduous, and examples I’ve seen run to tens of pages and others are longer (it varies due to different questions being asked based on earlier answers).

In contrast, the [biobank application PDF](#) is public and three and half pages long

Decision makers can not take information into account that biobank never asks for - even if biobank’s decision making on receipt of an application were to be perfect, the application form does not *ask* many questions.

**12. Does the biobank application form satisfy all the obligations of biobank to a) their cohort, b) each organisation who provides data to biobank, and c) GPs?**

I have been told, possibly not correctly, that the Board is provided with copies of all applications.

**13. Are the Board satisfied with the application process they oversee?**

**14. Relating to the answers to the previous two questions, what did the Board know, and when did they know it?**

### **Biobank projects**

**15. Did Biobank change anything after [media coverage](#) regarding China in 2022?**

A decade after the first research projects, biobank's website says they have around 3489 projects of which "40" are "complete".

**16. Is the "complete" number credible? How is it defined?**

**17. How many of those 3500 have gone bankrupt, changed ownership, or "pivoted"? Is that number more or less than 40?**

**18. Have biobank kept all promises made to AGD/IGARD and predecessors?**

Biobank publishes a list of projects on their website, making available only the applicant and lay summary, alongside an application number. Those application numbers are not sequential, but have large gaps, which may or may not be a problem, or an artefact of the process somehow. We have turned [the project list into a spreadsheet here](#).

A full analysis of all users will take investigators some time, initial examples show why a list of latest projects presented to all biobank members via the app might be politically difficult.

We did a brief look at projects largely at random: the second was Babylon (bankrupt and being sold off for parts), the third seemed also to have pivoted (no obvious biobank connection), and the fourth seemed to be... superficially fine. With a success rate of 25%, this appears like a job for someone with access to US corporate registration records more able to recognise the signs of deception. Sometimes, those signs are pretty clear.

The first project we looked at was Flying Troika.

### **Flying Troika**

No information is provided by biobank on whether a project got some data in a TRE, or whether they got access to genetic samples, etc, because biobank never publishes that information.

FlyingTroika.com, registered to the applicant, describes “A pure AI driven research lab” offering “[Flying Troika Deep Learning / AI / Machine Learning Research and Solutions](#)”.

They say “We are a pure research lab/ consulting service provider working on the cutting edge Deep Learning Technologies and Algorithms a variety of industries: Pharma, Medical Imaging, Finance, Insurance, Health Care, High Tech, Retail, Manufacturing, Oil & Gas, Industrial, Auto, Construction Applications.”.

Their first example is “Deep Learning in Finance and Insurance”, and list “OFFICE LOCATIONS/ RESEARCH & DELIVERY TEAMS” in “New York, Chicago, Los Angeles, San Francisco, London, Edinburgh, Hong Kong, Singapore, Miami, Paris, Berlin, Madrid, Barcelona”. 13 offices in 8 countries.

At the bottom of the webpage, are the customary twitter, facebook, and github logos, with a phone number given as “AI agent”.

Assuming biobank ever checked the webpage, they may have noticed a couple of typos in cities and the text, but what’s “Edinburg” or “Miami” amongst fellow researchers... The other business registered to the applicant [offers](#) subscriptions to a monthly supply of stock photos.

Had Biobank checked any other databases, they’d see that the company employs only two people, with a registered office at the same address as the domain is registered, which is a house in California.

- 19. Is this applicant real? What information did biobank receive, and what did biobank not ask for that they now should have?**
- 20. Would Flying Troika have passed NHS England’s processes?**
  - a. Do Flying Troika satisfy any obligations biobank has to NHS England if NHS England data was involved?**
- 21. What diligence was done on whether biobank or NHS data was used for “Finance” or “Insurance”? Is this what biobank participants signed up for?**
- 22. Does this project satisfy all assurances that GPs were given in the letter about what biobank does to ensure good governance?**
- 23. Was this organisation restricted to a TRE, or did someone send medical records and potentially genetic samples to a house in California?**
  - a. How does this relate to HDRUK’s strident objections to TREs for cohort studies?**

**24. How can biobank know the answers to this section if they didn't ask in the application process?**

**25. How would biobank PPIE and other engagement groups feel if this was the first example of a new project listed when they saw a new biobank tab appear in the NHS app?**

We expect a range of questions to apply for other biobank users; some of which will be similar, specifics will be different.

### **Areas of research with biobank**

The expansion of biobank registrations in recent years seems to have been largely driven by two overlapping groups, firstly are Artificial Intelligence and Machine Learning startups, and secondly are the emerging wave of "longevity" startups.

The argument seems to be that living forever reduces deaths, so the biobank team give access.

**26. Are there other companies like Flying Troika in the list of AI/ML/longevity startups that biobank has accepted?**

**27. Is this what cohort study PPIE/patient groups think that participants signed up for?**

Due to the nature of the "longevity" industry and the lack of checks, it is likely that newsworthy entities have been given access to biobank data, and biobank has not yet disclosed what they got.

One long standing critique, even of care.data, wherever coded data is collected there are the researchers arguing that they would like free text too. Some areas of care, such as mental health, are almost exclusively free text based. It is therefore unsurprising that some datasets have free text, either via MSHDS or via what biobank termed "intermediaries".

**28. What free text do relevant cohort studies have, and what are the data flows by which they get it?<sup>4</sup>**

We do not expect any one process to be perfect, but we expect every actor to act with integrity when something untoward happens, which is why stakeholders having a permanent seat at the decision making table is important, even if it is annoying to studies.

**29. As cohort studies wish access to GP data, will they produce a full audit of all data flows and data governance?**

---

<sup>4</sup> Have there been any accidents with MSHDS recently that NHS England would like to disclose?

**30. Can decisions be reviewed to the satisfaction of all the different constituencies of both BMA and AoMRC, as well as non-medical equivalencies?**

**31. How do cohort studies that started at birth deal with gillick competence and change over time about detailed records such as MH free text?**

## **Rest of Government**

Sir Ian Diamond's evidence to the Covid Inquiry relied upon his good chaps theory of access, as epitomised by the biobank, and this paper shows how that fails. Taking the data and abdicating governance because someone is seen as "trustworthy" today is a rapid route to them becoming the scandal of tomorrow. ONS treating admin data as akin to a survey is untenable over time, even without the text of the HDR Review by Cathie Sudlow being public.

As Sir Ian Diamond previously put it:

"There's no god given right for us to have data. There needs to be a really sound public good reason for collecting data, and using data, and people need to feel absolutely comfortable that their data are being used properly and kept securely and in a way that satisfies all forms of privacy"

– Sir Ian Diamond, National Statistician (in 2021<sup>5</sup>)

## **Summary: Sustainability means all stakeholders have a permanent seat at the table**

Biobank makes claims unsupported by their actions, and delivers insufficient transparency for the public to have confidence in their actions. Biobank's mishandling of GP communications, and the fact that biobank disseminates GP data already that they say GPs are "blocking" them having, needs additional clarity and likely process reform.

If HDR are correct that biobank are representative of all UKRI cohort studies, then all UKRI cohort studies may wish to check their existing data flows and practices as biobank will have to. We do agree, however, that whatever solution is acceptable and adopted should apply to *all* comparable UKRI cohort studies, as well as comparable others.

Due to the richness and sensitivity of GP data, the end state solution must require review of each project and the variables requested, so that every project that receives GP data has had the opportunity for review by the profession, as continues for the pandemic dataset via AGD/PAG.

Similarly, when OFH or ONS want access in to data bulk from DWP, DfE, MoJ, migration, or clubcard, (etc), each data controller should require individual project review for the project and the variables it wants.

---

<sup>5</sup> Sir Ian Diamond speaking at the Institute for Government <https://www.youtube.com/watch?v=tz9NihCTzqA&t=2373s> . It's unclear how this perspective has evolved towards the answers given to the committee.

Cohort studies should not be permitted to create their own data lake of GP data matching it back across projects, even if that may undermine the business case of some SDEs as a requirement of HDR's developments.

If a project applies for access to cohort data, it should say what data it wants, that access is approved by the cohort study, then approved by the relevant data controllers, and then accessed solely in a Good TRE sufficient for that data/study, wherever possible. Much of the application process can be invisible to the researcher as the questions should be the same.

Everything short of per project review will lead to more catastrophes with public trust, even if doing so project review is difficult today. It is the same with TREs. No cohort study who respects their cohort will argue, in the long term, that dangerous data practices are something they should continue to allow. The long running birth cohorts have a bond of trust with their participants that they see as akin to sacred, and TRE use will become part of that, just as computers became part of the way analysis worked and some entrenched academics argued they'd never use them.

If the Flying Troika and friends turn out to be akin to the actuaries working for insurers 2014, or the cancer registry's "causes of cancer study" run by a tobacco company, we would expect that the precedent of accountability and governance to result in leadership change. We have no expectation that biobank will take such responsibility, which is why they can not be trusted to make their own decisions without external scrutiny.

What allows programmes to survive is every stakeholder having a seat at every relevant table, and ensuring that promises are kept and actions are legitimate. Biobank's "trust us, we're biobank" has degraded badly. Ongoing project review and variable level access allows every GP to have confidence in biobank because JGPTIC was looking out for their wishes via AGD/PAG or equivalents.

medConfidential