Rapid projects support government departments to understand the scientific evidence underpinning a policy issue or area by convening academic, industry and government experts at a single roundtable. These summary meeting notes seek to provide accessible science advice for policymakers. They represent the combined views of roundtable participants at the time of the discussion and are not statements of government policy.

"How will human genome databases change between now and 2030, and who will hold them in the UK?"
Meeting notes from roundtable chaired by Prof John Iredale, facilitated by the Government Office for Science.
7th November 2023, 10:00-11:30

Key points:

- There will be an increase in the size and complexity of genetic datasets between now and 2030. The cost of computational power required to handle this data will also increase.
- There is a risk that the UK will not be able to maximise research outputs and health outcomes if new databases are set up in a siloed way and with existing databases.
- Genome databases must include diverse cohorts to ensure equity of research outcomes across different populations.
- There has been increasing use of AI with genetic technologies in the private sector. Use of such technologies could present opportunities or risks, and it is uncertain if companies can currently provide actionable insights.

There is a glossary at the end of this note covering technical terms.

1. How could primary and secondary genomic datasets change in the short and medium term?

- Primary datasets will increase in size and complexity due to increasing digitisation, the inclusion of "omics" and linked health data.
- Secondary datasets will develop over the next few years, including polygenic and integrated risk scores across large cohorts.
- Proliferation of these databases will result in data federation, where there is siloing of data resources, which could impact interoperability and limit researchers' ability to unleash the potential of UK based large scale cohorts when compared to data assets being held in fewer, larger databases.
- This problem could be compounded if non-clinical datasets are integrated with genetic data. While this siloing can in part be mitigated by enhancing interoperability, this has proved challenging to deliver technically and at a governance level. More effort seems to be going into building new secure data environments (SDEs) than in interoperability to join up existing databases or hosting existing resources.
- New Secure data environments (SDEs) should be designed or justified with specific purposes in mind, considering economies of scale for SDEs, research user needs, and ensuring patient data are used safely and for the public good (DHSC, 2022). As data and analyses become more complex, standards for metadata should be considered to give information about how genetic data have been collected, collated, or curated. Such specifications have been

developed by HDRUK, where data should be Findable, Accessible, Interoperable and Reuseable (FAIR) (UK Health Data Research Alliance, 2021; GO FAIR, 2016).

2. How could diversity of datasets change in the medium term?

- The need for greater diversity of datasets (and greater transparency in how we report datasets) is widely recognised in the UK (STANDING Together, 2023).
- Datasets should aim to over-represent minority populations so that research cohorts are large enough to ensure that such groups are not disadvantaged due to studies being underpowered (Ibrahim et al., 2021). Diverse datasets ensure particular health conditions can be explored equitably across the diversity of the population, and mean research can identify health conditions that may selectively affect groups within the population. This may also have value in predicting future healthcare demand, reflecting changing demographics of the population as a whole. To improve diversity of datasets, new ways of recruiting under-represented groups in the UK need to be considered, such as accessing newborn data from routine heel prick tests for research purposes (Cunningham-Burley et al., 2022).
- Low and middle income countries (LMICs) are producing their own datasets with global data on ancestry (International Health Cohorts Consortium, 2024), but there are concerns about exploitation of genetic information for commercial means, often at the expense of the communities from which data originates. This can also lead to a loss of trust in legitimate, equitable data collection initiatives.
- Sequence data will diversify due to new technologies in sequencing DNA as “long reads” rather than “short reads”. Long read sequencing gives more information about DNA structure and function than short reads. There is uncertainty whether the shift towards long read sequencing will happen in the next five years.

3. How could analyses of health and lifestyle insights from genomic datasets change?

- There is an opportunity to link genetic data with wider information. Beyond linking clinical records, there is potential to include links to non-clinical data such as social and educational data. Links to non-clinical data are already happening in the UK: for example, consent options for research participants for the Generation Scotland cohort includes linkage to education outcomes and tax records. More interaction between the health system and researchers through non-clinical data could improve our understanding of health determinants.
- In addition to DNA sequencing, technological advances will also enable large-scale generation of other types of molecular data, such as epigenetics and proteomics. Such datasets have the potential to provide actionable information as they more closely reflect an individual’s current health status.
- The design of SDEs needs to ensure that, in addition to their role as ‘data banks’, they also have the analytical tools and computational power required for cutting-edge reproducible research and innovation. Many are adopting full cloud or hybrid approaches (as opposed to full on-premises solutions) to enable scale-up and scale-down according to researcher need (NHS, 2023). However, all models of computation at this level are costly, and there is still uncertainty as to the financial models that ensure that we simultaneously achieve financial sustainability for SDEs, affordability for researchers, and provision of adequate tools and computational power to researchers.

4. What anticipated uses of genomic data could we expect to see in the future that are currently unfeasible?

- A range of new short read technologies could facilitate huge drops in sequencing prices. One US company has reported $100 per genome (Pennisi, 2022), but the group felt it was likely this will be an order of magnitude cheaper by 2030. At this lower price point, it was suggested that whole population genome sequencing could be feasible.
- There has been a large increase in private companies using AI. A recent report found AI and genomics becoming increasingly intertwined, with the market for AI and genetic technologies reaching £19.5bn by 2030, up from £0.5bn in 2021. It found both opportunities and risks from using AI, such as biases being introduced from training on unrepresentative or biased data (Nuffield Council on Bioethics, 2023). There are several factors shaping the genetic technology sector: advances in technology, reduction in cost and commercial factors, public perception, and regulation and policy.
- Increasingly, private companies sell genomic products on the basis of a claim that they support personalised medicine in line with an individual's genome. However, it is uncertain if companies can deliver accurate or actionable insights to consumers, or if the UK healthcare system is currently able to meaningfully respond to these sources of information about patients.

5. What are the implications of current ownership of genome databases in the UK?

- It is unclear whether there is a trusted environment system for long-term archiving and sharing of data from cohorts when they are 'closed', such as when funding ends.
- Smaller genome databases arising from specific research cohorts may still be valuable to the wider research community but be under-used due to not being findable or appropriately accessible. All future cohorts regardless of size should ascribe to the FAIR principles, and funders and data hosting institutions should consider how they can operationalise these principles (GO FAIR, 2016).
- There was consensus that genetic data should be seen as part of standard NHS data. There is ongoing debate around whether researchers should be charged for use of NHS data (NHS, 2023). A risk is that commercialisation could lead to reduction in data use for research, but uncertainty on what the charging model will be is having a detrimental impact on long-term planning of research. The Biobank model, where researchers pay to access data and use the research analysis platform, could be replicable across the public sector.

6. What could database ownership look like in 2030 across different sectors?

- 2030 could see increasing numbers of entirely private datasets, which are catching up with public datasets in terms of amount of data held. This could include direct-to-consumer test companies, but also pharmaceutical and broader life science companies. Private datasets could have huge research value, but companies have commercial incentive for them to remain exclusive.
- UK Biobank could be a model for pre-competitive collaborative funding of research databases, where a private sector (or public-private) consortium funds additional data generation (e.g. UK Biobank whole-genome sequencing via public and pharma partners) and holds data exclusivity over a short period, before the data becomes widely available. This model addresses concerns over data ownership as UKB owns the data once it is incorporated into the resource while unlocking private investment. In some models, private companies could enrich existing public datasets by contributing it to a central data environment. Having a simple, transparent, and equitable access policy for all researchers (i.e. academic and commercial) with no differential access fees has contributed to this successful model, as the commercial sector has made subsequent investments to enhance the resource.

- Our Future Health is a programme funded by a collaboration between the public, charity, and private sectors. It is approaching ownership as the 'custodian' of participants' data, and has incorporated participants, funders, stakeholders, and the public in its governance model; with an emphasis on collaboratively designed policies and procedures to help guide decisions about data uses.
- Ownership of databases collected by private companies can change if databases are sold, including internationally (Lillington, 2023). There is a risk that perceived (or actual) breaches of public trust through such sales could impact public interactions with genomic data, including making people less likely to donate genomes and impacting future cohorts.
- SMEs are lower profile and therefore represent fewer risks to public perception. They are more likely to be spin outs from universities and deliver a specific product or output rather than general analysis. Timely and commercially viable access to data is important to SMEs, as many have relatively short-term funding and need to rapidly demonstrate profitability.

7. What are the possible opportunities or ramifications of any shift in ownership?

- There could be an opportunity to develop good public "social licence to operate" contracts with participants, encouraging participation in cohort studies by demonstrating flow from public datasets to public good, such as development of new diagnostics and therapies.
- A potential risk is in areas of genomics where data is mostly held in private companies, particularly in cancer. Private companies are the largest repositories of cancer genomics and keep data internally. There is a risk that this will cause progress in personalised medicine for cancer to slow, owing to fewer researchers having access to data.
- There could be improved interaction between patients and their data, such as feedback to participants about health-related insights generated from research. There is increased interest among patients in accessing their own data.

Participants:

John Iredale (Chair, University of Bristol), Alastair Denniston (Regulatory Horizons Council (RHC) & Alan Turing Institute), Andy Green (Regulatory Horizons Council (RHC) & University of Oxford), Arzoo Ahmed (Our Future Health), Catalina Vallejos (University of Edinburgh), Clive Darwell (UK Biobank), Matt Brown (Genomics England), Olivier Roth (BioIndustry Association), Riccardo Marioni (University of Edinburgh & Generation Scotland Cohort), Sophia McCully (Nuffield Council on Bioethics), Thomas Keane (EMBL-EBI), Tim Hubbard (KCL, WHO Therapeutic Advisory Group for Genomics & HDR-UK)

Secretariat: Government Office for Science Officials
Observers: Office for Life Sciences Officials

Glossary

Epigenomics/epigenetics: Studies of how biological regulation of DNA (for example due to environmental factors), rather than DNA sequence, affects the way genes work (CDC, 2022).

Integrated risk scores: A measure of risk of developing a specific disease taking into account an individual's polygenic risk score and other factors such as age and lifestyle (Our Future Health, 2022).

Omics: Studies aiming to measure biomolecules that translate into structure, function, and dynamics of an organism. The suffix "-ome" is used to describe the totality of a specific molecule being studied, for example "genome" refers to all the DNA in an organism (Richards, 2021).

Polygenic Risk Scores: A measure of risk of developing a specific disease due to an individual's genetic variation (Our Future Health, 2022).

Primary datasets: Databases that contain direct sequence information and annotations from experimentally derived data (EMBL, 2023).

Proteomics: Studies of the entire complement of proteins in an organism (Richards, 2021).

Secondary datasets: Databases that summarise analyses of DNA sequence information (EMBL, 2023).

Secure data environments (SDEs): Online platforms that give approved users access to health data for analysis in a secure environment (Transformation Directorate, 2023). These are sometimes referred to by other names like "trusted research environments".

References:

CDC. (2022). What is Epigenetics? Available at: https://www.cdc.gov/genomics/disease/epigenetics.htm (accessed 27/02/2024)

Cunningham-Burley, S. et al. (2022). Feasibility and ethics of using data from the Scottish newborn blood spot archive for research. Commun Med 2, 126 (2022). https://doi.org/10.1038/s43856-022-00189-2

DHSC. (2022). 'Data saves lives: reshaping health and social care with data' Available at: https://www.gov.uk/government/publications/data-saves-lives-reshaping-health-and-social-care-with-data/data-saves-lives-reshaping-health-and-social-care-with-data#improving-trust-in-the-health-and-care-systems-use-of-data (accessed 27/02/2024)

EMBL. (2023). Primary and secondary databases. Bioinformatics for the terrified. Available at: https://www.ebi.ac.uk/training/online/courses/bioinformatics-terrified/what-makes-a-good-bioinformatics-database/primary-and-secondary-databases/ (accessed 27/02/2024)

GO FAIR. (2016). 'FAIR Principles', Available at: https://www.go-fair.org/fair-principles/ (accessed 27/02/2024)

Ibrahim, H. et al. (2021). Health data poverty: an assailable barrier to equitable digital health care. The Lancet Digital Health, Volume 3, Issue 4. DOI: https://doi.org/10.1016/S2589-7500(20)30317-4

International Health Cohorts Consortium. (2024). IHCC Cohort Atlas, Available at: https://atlas.ihccglobal.org/ (accessed 27/02/2024)

Lillington, K. (2023) Future of Genuity's Irish DNA database is worryingly uncertain, Irish Times, Available at: https://www.irishtimes.com/technology/2023/09/07/future-of-genuitys-irish-dna-database-is-worryingly-uncertain/ (accessed 27/02/2024)

NHS. (2023). Value Sharing Framework for NHS data partnerships, Available at: https://transform.england.nhs.uk/key-tools-and-info/centre-improving-data-collaboration/value-sharing-framework-for-nhs-data-partnerships/ (accessed 27/02/2024)

Nuffield Council on Bioethics. (2023). AI and genomics futures. Available at: https://www.nuffieldbioethics.org/publications/ai-and-genomics-futures (accessed 27/02/2024)

Our Future Health (2022) Genomics plc to generate polygenic risk scores for Our Future Health. Available at: https://ourfuturehealth.org.uk/news/genomics-plc-to-generate-polygenic-risk-scores-for-our-future-health/ (accessed 27/02/2024)

Pennisi, E. (2022). A $100 genome? New DNA sequencers could be a 'game changer' for biology, medicine. Science. Available at: https://www.science.org/content/article/100-genome-new-dna-sequencers-could-be-game-changer-biology-medicine (accessed 27/02/2024)

Richards, S. (2021). Omics made easier. FRED HUTCH NEWS SERVICE. Available at: https://www.fredhutch.org/en/news/center-news/2021/03/omes-omics-primer.html (accessed 27/02/2024)

STANDING Together. (2023). 'About: STANDING Together: Building STANdards for data Diversity, INclusivity, & Generalisablity' Available at: https://www.datadiversity.org/about (accessed 27/02/2024)

Transformation Directorate. (2023). Secure Data Environments (SDEs). NHS. Available at: https://transform.england.nhs.uk/key-tools-and-info/data-saves-lives/secure-data-environments/ (accessed 27/02/2024)

UK Health Data Research Alliance. (2021). 'Recommendations for Data Standards in Health Data Research White Paper', Available at: https://ukhealthdata.org/wp-content/uploads/2021/12/211124-White-Paper-Recommendations-of-Data-Standards-v2-1.pdf (accessed 27/02/2024)