UK health data research infrastructure must take privacy risk seriously

Luc Rocher[1], Jessica Morley,[2]

[1.] Oxford Internet Institute, University of Oxford, Oxford, UK

[2.] Digital Ethics Center, Yale University, USA

Correspondence to: J Morley jessica.morley@yale.edu

Leaks from the UK Biobank undermine trust and willingness to share health data

UK Biobank is one of the world's most important biomedical research resources, holding detailed genetic, health, and lifestyle data on half a million British volunteers. More than 22 000 researchers across 60 countries have used the data, contributing to over 18 000 peer reviewed publications.[1] It is an extraordinary resource, with a tarnished reputation after a series of scandals In 2024, a team promoting racist pseudoscience was filmed claiming to have obtained Biobank data.[2] In 2025, intelligence officials expressed concerns after one in five successful access applications came from China.[3] Academics have argued Biobank's consent documents and leaflets have left participants in the dark about these risks.[4]

Most recently, an investigation by the *Guardian* newspaper revealed that researchers have—in the past—repeatedly and accidentally uploaded datasets to the code sharing platform GitHub, forcing Biobank to issue at least 80 legal takedown notices. One file included millions of hospital diagnoses and appointment dates for over 400 000 participants.[5]

UK Biobank has sought to downplay the importance of the exposure, stating that the data posted to GitHub are "de-identified" and free of "personally identifying information," and that no participant has been unwillingly re-identified. But the science is clear: in datasets containing common demographic and health attributes, individual records are often unique and vulnerable to re-identification, even when datasets are incomplete.[6] The *Guardian* showed this by correctly identifying a single participant from just two easily known facts.[5] This is precisely why UK law does not rely on the concept of de-identification to determine the level of protection required. Guidance from the Information Commissioner's Office regards data that could be re-identified using publicly available resources and standard investigative techniques as personal data.[7]

Acknowledging the true risk of re-identification matters. When institutions treat privacy as a box ticking exercise, the public often pushes back. In England, a planned centralised database of anonymised general practice records, Care.data, collapsed in 2016 after privacy concerns resulted in 1.5 million people opting out of their data being used for purposes beyond direct care.8 The General Practice Data for Planning and Research (GPDPR) programme had to be delayed in 2021 after a similar backlash saw opt-outs almost double in a single month.[9] The current rate of opt outs is 5.58% of the population of England.

Lack of trust threatens health

The government is betting heavily on health data and artificial intelligence in its 10 year health plan for England, which commits to making the NHS "the most AI-enabled health system in the world."10 Over £700m of public and charitable investment is flowing into the National Data Lib-

rary and Health Data Research Service to support this vision.[11] [12] Were the public to conclude that the government and its research partners are cavalier about the risks, data might stop flowing. Given that the AI opportunities action plan names Biobank as a resource that the National Data Library should build upon,[13] and that in February 2026 the government paved the way for linkage between general practice records and Biobank to serve precisely this purpose,[14] Biobank's lacklustre response to the latest *Guardian* revelations could inflict on these ambitions is considerable.

If the UK cannot provide datasets that are both representative and comprehensive, developers will look elsewhere. Britain could then become increasingly reliant on AI models trained overseas rather than developed domestically. Not only would this undermine the ambitions of the AI opportunities action plan,[15] but it would put patients at risk. Risk stratification tools are, for example, already routinely deployed far outside the context in which they were developed, often without adequate international validation, and evidence suggests this translates into poor performance and, in some cases, measurable harm.16 Underserved communities—those with the strongest historical reasons to distrust the health system and therefore the most likely to withdraw their data when trust is damaged—will likely be the most harmed by this pattern. Minority ethnic groups, for instance, are already among the least willing to share their data for research purposes,[17] and the most likely to be harmed by poorly targeted AI models.[15] For a healthcare system premised on the values of equity and solidarity, this should be an unpalatable consequence.

This can be avoided if all players in the health data research ecosystem take privacy and security seriously by investing in sociotechnical infrastructure that offers better privacy protection while still maximising utility. There are two aspects to this. The first is technical: taking a defence-in-depth approach with trusted research environments, layered protections, and robust governance while making that architecture transparent. This way, participants know how their data are being used and by whom. Biobank's 2024 decision to end direct downloads and require all analyses to take place within its research analysis platform was a step in this direction.

The second is attitudinal. As well as mitigating technical risks, institutions must demonstrate humility, a commitment to listening to privacy experts, and a willingness to learn. Just as the NHS monitors patient safety, the aspiration should be to reduce risks as much as possible while accepting that no system is perfectly secure, that participants deserve to know when something goes wrong, and that incidents must be investigated, not to "manage" them but to learn lessons. It is time the entire research community took this responsibility seriously.

References

1 UK Biobank. Use our data. 2025. https://www.ukbiobank.ac.uk/use-our-data/

2 Pegg D, Devlin H, Burgis T. Concerns raised over access to UK Biobank data after 'race scientists' claims. Guardian. 25 Oct 2024. https://www.theguardian.com/science/2024/oct/25/concerns-raised-access-uk-biobank-data-race-scientists-claims

3 Burgis T. Revealed: Chinese researchers can access half a million UK GP records. Guardian. 15 Apr 2025.

4 Barn G. Consent and its discontents: the case of UK Biobank. *Med Health Care Philos* 2025;28:533-47. doi:10.1007/s11019-025-10276-5. PubMed

5 Devlin H, Burgis T, Hoyland L. Confidential health records from UK BioBank project exposed online. Guardian. 14 Mar 2026. https://www.theguardian.com/science/2026/mar/14/confidential-health-records-exposed-online-uk-biobank

6 Rocher L, Hendrickx JM, de Montjoye Y-A. Estimating the success of re-identifications in incomplete datasets using generative models. *Nat Commun* 2019;10:3069. doi:10.1038/s41467-019-10933-3. PubMed

7 Information Commissioner's Office. How do we ensure anonymisation is effective? 2025. https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/data-sharing/anonymisation/how-do-we-ensure-anonymisation-is-effective/

8 Limb M. Controversial database of medical records is scrapped over security concerns. *BMJ* 2016;354:i3804. doi:10.1136/bmj.i3804. PubMed

9 Wise J. Government tried to launch patient data scheme without right safeguards, MPs are told. *BMJ* 2022;377:o1208. doi:10.1136/bmj.o1208. PubMed

10 Department of Health and Social Care, Prime Minister's Office, 10 Downing Street. Fit for the future: 10 Year Health Plan for England. 2025.

11 Department for Science. Innovation and Technology. National Data Library: progress update, January 2026. 2026. https://www.gov.uk/government/publications/national-data-library-progress-update-january-2026/national-data-library-progress-update-january-2026

12 Wellcome Trust. National data service will simplify access to health data for research. 2025. https://wellcome.org/insights/articles/national-data-service-will-simplify-access-health-data-research

13 Department for Science. Innovation and Technology. AI Opportunities Action Plan. 2025. https://www.gov.uk/government/publications/ai-opportunities-action-plan/ai-opportunities-action-plan

14 Fiddy E. Major milestone for health research as UK Government decision enables access to UK Biobank volunteers' GP patient data. UK Biobank. 2026. https://www.ukbiobank.ac.uk/news/major-milestone-for-health-research-as-uk-government-decision-enables-access-to-uk-biobank-volunteers-gp-patient-data/

15 Prime Minister’s Office. PM speech on AI opportunities action plan: 13 January 2025. 2025. https://www.gov.uk/government/speeches/pm-speech-on-ai-opportunities-action-plan-13-january-2025

16 Oddy C, Zhang J, Morley J, Ashrafian H. Promising algorithms to perilous applications: a systematic review of risk stratification tools for predicting healthcare utilisation. *BMJ Health Care Inform* 2024;31:e101065. doi:10.1136/bmjhci-2024-101065. PubMed

17 Jones LA, Nelder JR, Fryer JM, et al. Public opinion on sharing data from health services for clinical and research purposes without explicit consent: an anonymous online survey in the UK. *BMJ Open* 2022;12:e057579. doi:10.1136/bmjopen-2021-057579. PubMed