Comment

CARDIAC RESEARCH

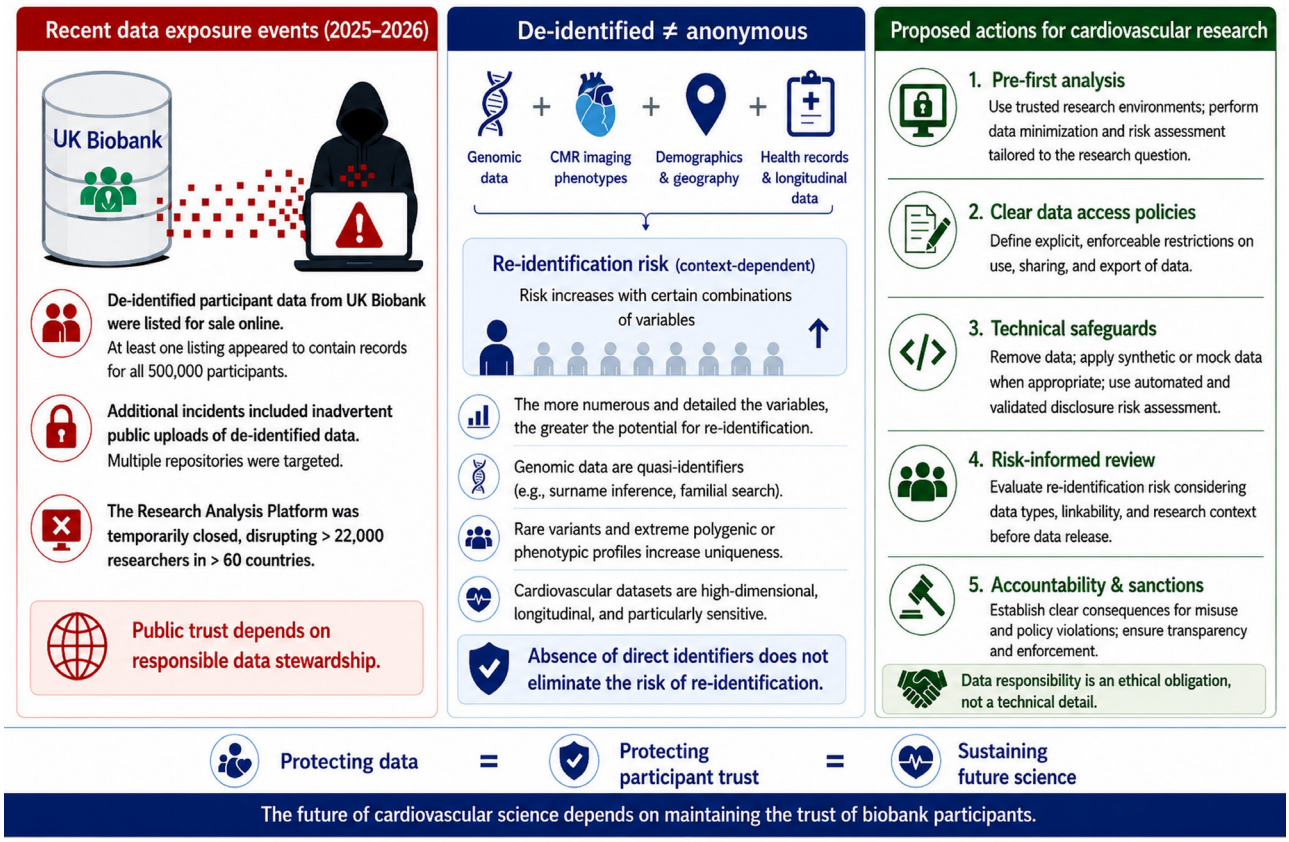
OPEN

Safeguarding the world's biobanks

A call to cardiovascular science

Yang Liu, PhD[a,*]

Graphical abstract



In April 2026, UK Biobank reported that de-identified participant data released to approved investigators under contractually binding access agreements had appeared for sale on a consumer e-commerce platform.[1] According to UK Biobank, the affected data had been provided to investigators at a small number of academic institutions whose access has been suspended pending ongoing investigation. The UK Government stated to the House of Commons on April 23 that at least one listing was reported to contain records from all 500,000 participants.[2] The exposed variables reportedly included sex, year and month of birth, socioeconomic indicators, lifestyle factors, and biological-sample-derived measurements, but not direct identifiers such as names, addresses, or National Health

Supplemental Digital Content is available for this article.

[a] Department of Epidemiology and Biostatistics, School of Public Health, Peking University, Beijing, China.

* Correspondence: Yang Liu, Department of Epidemiology and Biostatistics, School of Public Health, Peking University, 38 Xueyuan Road, Beijing 100191, China (e-mail: yang_liu@bjmu.edu.cn).

© The Author(s) 2026. Published by Wolters Kluwer Health, LLC. on behalf of Higher Education Press. This is an open access article distributed under the Creative Commons Attribution License 4.0 (CCBY), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Cardiac Research (2026) 2:2

Received: 28 April 2026 / Received in final form: 6 May 2026 / Accepted: 9 May 2026

http://dx.doi.org/10.1097/re9.0000000000000017

Service (NHS) numbers.[1,2] Although the listings were removed before any sale was completed, the incident led UK Biobank to close its Research Analysis Platform and introduce new export controls, interrupting active analyses by more than 22,000 approved researchers in over 60 countries.[1]

Nor was this an isolated incident. In March 2026, UK Biobank acknowledged that researchers had inadvertently uploaded portions of de-identified participant data to public code repositories alongside analytical scripts.[3] An independent tracker maintained by the Oxford Internet Institute has documented approximately 110 GitHub copyright takedown notices targeting around 197 repositories across at least 14 countries between July 2025 and April 2026, though these figures have not been independently audited.[4] The UK Parliament's Science, Innovation and Technology Committee has expressed concern that lessons may not have been learned.[5] For cardiovascular science, one of the fields most deeply shaped by the UK Biobank, these events are not a UK problem. They are a global warning.

1. Cardiovascular science's debt to biobank participants

Modern cardiovascular epidemiology has been transformed by population-scale resources that the Framingham generation could scarcely have imagined.[6] UK Biobank, with genome-wide genotyping in approximately 488,000 participants, cardiovascular magnetic resonance (CMR) imaging at an unprecedented scale, and linkage to national health records, has become one of the most influential platforms in contemporary cardiovascular research.[7] Polygenic scores for coronary artery disease developed in the UK Biobank have identified subgroups with risks comparable to monogenic familial hypercholesterolemia, reshaping primary prevention;[8] CMR-derived endophenotypes have uncovered novel loci for left ventricular structure and enabled polygenic prediction of dilated cardiomyopathy.[9] The cardiovascular community is therefore among the clearest beneficiaries of the trust placed in UK Biobank by its 500,000 volunteers. That privilege cannot be discharged by a methods-section citation alone.

2. "De-identified" is not "anonymous"

Public communications about the recent incident emphasized that the exposed data did not contain personally identifying information such as names, addresses, exact dates of birth, or NHS numbers, while acknowledging that the possibility of identification can never be completely removed.[1,2] Such data should not be treated as anonymous merely because direct identifiers are absent. In regulatory and operational terms, they are better understood as de-identified or pseudonymized data whose residual identifiability depends on context, linkage, and access conditions. Under both UK and EU law, pseudonymized data remain personal data precisely because reidentification, while restricted, is not eliminated. Rocher and colleagues[10] estimated that 99.98% of Americans could be correctly reidentified using 15 demographic attributes, and genomic data also function as quasi-identifiers through surname inference[11] and familial searches in genealogy databases.[12] For cardiovascular research, these risks are especially salient: CMR-derived phenotypes provide high-dimensional descriptors of individual biology; carriers of rare cardiomyopathy variants (e.g., truncating Titin or pathogenic the cardiac isoform of Myosin Binding Protein C 3) belong to small subgroups; and individuals at the extremes of polygenic risk distributions are similarly uncommon. When combined with month-and-year of birth, geographic indicators, and longitudinal hospital data, the cumulative reidentification risk may be greater than the absence of direct identifiers suggests.

3. A global, not a British, problem

UK Biobank is the most visible, but not the only, example of a new generation of population-scale resources, alongside the All of Us Research Program, the China Kadoorie Biobank, the Kailuan Study, BioBank Japan, FinnGen, the German National Cohort, and Our Future Health. Each has made cardiovascular contributions UK Biobank alone could not provide: All of Us, BioBank Japan, and the China Kadoorie Biobank have enabled trans-ancestry validation of polygenic risk scores for coronary artery disease and atrial fibrillation;[13] The China Kadoorie Biobank and Kailuan Study have generated indispensable evidence on East Asian-specific risk factors; FinnGen has identified rare loss-of-function variants relevant to lipid metabolism and heart failure; and the German National Cohort and Our Future Health are extending deep cardiovascular phenotyping at multimillion-participant scale. Collectively, these resources are indispensable for international validation of cardiovascular risk prediction tools such as Systematic Coronary Risk Evaluation 2 and the Pooled Cohort Equations beyond their derivation populations. A major data incident in one prominent resource risks eroding confidence in others and slowing efforts to make cardiovascular prevention equitable across populations.

4. The downstream consequences of inaction

Beyond immediate harm, the 2026 incidents could generate three downstream consequences for cardiovascular research. First, custodian-level tightening of data export controls may impede cross-cohort external validation of cardiovascular risk algorithms, encouraging within-cohort estimates where rigorous external validation is needed. Second, cardiovascular artificial intelligence (AI) models, including deep-learning electrocardiogram (ECG) and CMR phenotyping pipelines,[14] depend on the secure flow of large datasets across institutions;

restrictive environments without adequate computational infrastructure could slow improvements in model generalizability. Third, repeated data-governance failures risk eroding public trust;[15] future longitudinal studies may struggle to recruit participants if the perceived risk of a data incident outweighs the perceived benefit. Each consequence strengthens the case for action: restrictive environments must be paired with adequate computational capacity, federated analysis frameworks,[16] and harmonized cross-jurisdictional access procedures, so legitimate international validation does not become collateral damage of legitimate security tightening.

5. A call to cardiovascular science

Editors of cardiovascular journals are not data custodians, regulators, or forensic auditors. Primary responsibility for biobank governance rests with custodians, ethics-review bodies, institutions, and code-hosting platforms. Journals nevertheless have a narrow but important point of leverage: they can align open-science requirements with the contractual and ethical obligations that govern population-scale participant data. First, journals should require a trusted research environment-default analysis statement, declaring whether individual-level analyses, including CMR-derived endophenotyping, deep-learning ECG interpretation, and polygenic risk score construction, were conducted within a custodian-approved secure platform.[17] A model statement is provided in Supplementary Box 1, Supplemental Digital Content, https://links.lww.com/CARES/A13. Second, journals should replace generic availability statements with biobank-specific language. Generic language such as "data available upon reasonable request" is incompatible with most biobank access agreements; an appropriate model statement is provided in Supplementary Box 2, Supplemental Digital Content, https://links.lww.com/CARES/A13. Third, journals should require safeguards for linked code repositories. In cardiovascular research, this risk is amplified by ECG waveforms, CMR digital imaging and communications in medicine-derived arrays, and PLINK or BGEN genotype files that may be embedded in Jupyter or R notebooks. A 5-item submission checklist for repository self-certification is provided in Supplementary Box 3, Supplemental Digital Content, https://links.lww.com/CARES/A13. Fourth, journals should flag studies with high residual reidentification risk for additional governance review, aligned with established frameworks such as National Institute of Standards and Technology (NIST) Special Publication 800-188[18] and minimum-cell-size (k-anonymity) rules. Operational criteria are detailed in Supplementary Table 1, Supplemental Digital Content, https://links.lww.com/CARES/A13. Fifth, journals should help build accountability beyond the journal page. Journals and data custodians should establish incident-notification channels for suspected violations of data-use agreements; institutions should investigate promptly and transparently; sanctions should be proportionate. Compliance with biobank access agreements is not bureaucratic friction, but an ethical obligation to the participants whose trust made the research possible.

6. Beyond journals: complementary roles for institutions, funders, and researchers

Sustained progress requires complementary commitments from the wider ecosystem. Institutions should embed biobank-data governance into existing research-integrity infrastructure, with mandatory training on access agreements, output-checking rules, and reidentification risks, and should treat compliance with data-use agreements as equivalent to ethics-approval compliance. Funding agencies should require, as a condition of award, that grants specify the secure analytical environment to be used and budget for working within it. Individual researchers, especially computationally trained cardiovascular trainees, should receive structured training in biobank governance as a core component of methodological education, alongside study design and statistical inference.

7. The social contract we cannot afford to break

It would be a profound mistake to interpret these events as an argument against open science. Open science requires open methods, open code, and reproducible analysis, all compatible with strict custodianship of the underlying records. Trusted research environments have been criticized for seeking to replace interpersonal trust with technical guarantees;[19] yet the UK Biobank suspension illustrates that uncontrolled data export can disrupt science at a scale often attributed to secure-environment policies. The closure of UK Biobank's Research Analysis Platform interrupted active analyses by more than 22,000 approved researchers in over 60 countries, a reminder that weak governance can impede science as surely as excessive restriction.[1] Technical safeguards can protect participant trust; they cannot replace it. Half a million people in the UK, and millions more worldwide, agreed to share their lives with science because they were told it would help others. The task now is to ensure the infrastructure through which we work is worthy of the next generation, whose trust we will ask for.

Acknowledgements

Not applicable.

Ethical statements

No new human participant data were collected or analyzed; the article discusses publicly reported data-governance issues.

Conflicts of interest

The author has no conflicts of interest to disclose.

Funding source

Not applicable.

Data availability statement

Data sharing is not applicable to this article as no datasets were generated or analyzed during the current study.

Author contributions

Yang Liu wrote the manuscript, reviewed, and revised the proof.

AI statement

No generative AI or AI-assisted technologies were used in the preparation of this manuscript. The author takes full responsibility for the content.

References

[1] Collins SR. A message to our participants: UK Biobank data security update. UK Biobank; 23 April 2026. Available at: ukbiobank website. Accessed 2 May 2026.
[2] Murray I. Minister of State statement to the House of Commons: 23 April 2026. Department for Science, Innovation and Technology, UK Government; 23 April 2026. Available at: GOV.UK website. Accessed 2 May 2026.
[3] Collins SR. A message to our participants: protecting your personal information. UK Biobank; 14 March 2026. Available at: ukbiobank website. Accessed 2 May 2026.
[4] Rocher L. UK Biobank health data keeps ending up on GitHub. Oxford Internet Institute; April 25, 2026. Available at: ukbiobank website. Accessed 2 May 2026.
[5] Onwurah C. Biobank data leak: Science, Innovation and Technology Committee responds. UK Parliament; 23 April 2026. Available at: UK Parliament website. Accessed 2 May 2026.
[6] Sudlow C, Gallacher J, Allen N, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med. 2015;12:e1001779.
[7] Bycroft C, Freeman C, Petkova D, et al. The UK Biobank resource with deep phenotyping and genomic data. Nature. 2018;562:203–9.
[8] Khera AV, Chaffin M, Aragam KG, et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nat Genet. 2018;50:1219–24.
[9] Pirruccello JP, Bick A, Wang M, et al. Analysis of cardiac magnetic resonance imaging in 36,000 individuals yields genetic insights into dilated cardiomyopathy. Nat Commun. 2020;11:2254.
[10] Rocher L, Hendrickx JM, de Montjoye YA. Estimating the success of re-identifications in incomplete datasets using generative models. Nat Commun. 2019;10:3069.
[11] Gymrek M, McGuire AL, Golan D, Halperin E, Erlich Y. Identifying personal genomes by surname inference. Science. 2013;339:321–4.
[12] Erlich Y, Shor T, Pe'er I, Carmi S. Identity inference of genomic data using long-range familial searches. Science. 2018;362:690–4.
[13] Patel AP, Wang M, Ruan Y, et al.; Genes & Health Research Team; the Million Veteran Program. A multi-ancestry polygenic risk score improves risk prediction for coronary artery disease. Nat Med. 2023;29:1793–803.
[14] Attia ZI, Kapa S, Lopez-Jimenez F, et al. Screening for cardiac contractile dysfunction using an artificial intelligence-enabled electrocardiogram. Nat Med. 2019;25:70–4.
[15] Critchley C, Nicol D, Otlowski M. The impact of commercialisation and genetic data sharing arrangements on public trust and the intention to participate in biobank research. Public Health Genomics. 2015;18:160–72.
[16] Rieke N, Hancox J, Li W, et al. The future of digital health with federated learning. NPJ Digit Med. 2020;3:119.
[17] Goldacre B. Better, broader, safer: using health data for research and analysis. Department of Health and Social Care, UK Government; April 7, 2022. Available at: GOV.UK website. Accessed 2 May 2026.
[18] Garfinkel S, Guttman B, Near J, Dajani A, Singer P. De-identifying government datasets: techniques and governance. NIST Special Publication 800-188. National Institute of Standards and Technology; 2023. Available at: ukbiobank website. Accessed 2 May 2026.
[19] Graham M, Milne R, Fitzsimmons P, Sheehan M. Trust and the goldacre review: why trusted research environments are not about trust. J Med Ethics. 2023;49:670–3.