# Realising the full potential of Our Future Health through data linkage and trans-biobank efforts

Vincent J. Straub[1,*], Stefania Benonisdottir[1,2], Augustine Kong[1], and Melinda C. Mills[1,3,4,*]

[1] Leverhulme Centre for Demographic Science, Nuffield Department of Population Health, University of Oxford and Nuffield College, Oxford, UK
[2] Institute of Physical Sciences, University of Iceland, Reykjavík, Iceland
[3] Department of Genetics, University Medical Centre Groningen, Groningen, The Netherlands
[4] Department of Economics, Econometrics and Finance, University of Groningen, Groningen, The Netherlands

*Correspondence to: vincent.straub@ndph.ox.ac.uk; melinda.mills@demography.ox.ac.uk;

**The scale and population coverage of Our Future Health, alongside other next generation biobanks, offers unprecedented opportunities to advance genomic medicine. Focusing on the UK context, we provide a researcher perspective of how this new resource could reach its full potential in a way that is impactful, user-friendly, and informs related global efforts.**

Biobanks, entities that systematically collect, store and manage biological specimens, are reshaping medical research and global health by enabling unprecedented scale and depth[1]. In this Comment, we discuss how Our Future Health represents a major new biobank resource and consider how it can realise its full potential in advancing population-scale genomic research. We first highlight how its broad coverage of the UK adult population can counter past biases and set a new norm of multi-ancestry cohort studies. We then consider how it is uniquely positioned to catalyse data linkage and trans-biobank research, thereby complementing census or other regularly collected data to deepen our understanding of complex traits, including parent-of-origin effects. Finally, we consider how it can inform related international efforts, and end with highlighting key research priorities for Our Future Health to continue to shape the global genomics landscape.

## Building on the success of existing large-scale biobanks

The prevention, detection, and treatment of diseases has benefited enormously from multiple observational cohort studies, particularly large, so-called ‘mega-biobanks’ with data on >100K individuals. Pioneers include the UK Biobank (UKBB) and Million Veterans Programme (MVP) (**Fig. 1b**). These biobanks were distinct from previous efforts in that they were often national efforts including thousands of individuals, collected diverse biological samples (e.g., blood, saliva, urine, and DNA), offered deeper and comprehensive phenotyping (e.g., demographics,

lifestyle, family history), and were longitudinal or linked to health or other heath-related national registry data (e.g., tax, employment, education records).

Within the last few years, other large databases have emerged including FinnGen (https://www.finngen.fi/en) in 2017, the All of Us Research Program (https://allofus.nih.gov/) in 2018, and, most recently, Our Future Health (https://ourfuturehealth.org.uk/). The latter, which began recruitment in 2022, is a prospective, observational cohort study of the UK adult population that aims to recruit 5 million participants. Whilst it builds on the success of existing biobanks, several features set Our Future Health apart (**Box 1**)—chief among these: its scale. At the time of writing, it already holds data on over 1.7 million participants (**Fig. 1a**), making it the largest biomedical dataset in the world[2]. Given its target enrolment, it is set to recruit ~10% of the UK adult population. Whilst this is still below the population coverage of biobanks in the smaller northern European countries of Iceland (deCODE), Finland (FinnGen), and Estonia (EstBB), the sheer size of Our Future Health means it will be orders of magnitude larger (**Supplementary Table 1**).

## Committing to broader coverage and investigating bias

Many biobanks face the challenge of volunteer-based, convenience sampling, which often leads to cohorts that do not fully represent the target population. Our Future Health has attempted to address this issue by recruiting participants through clinics distributed across the UK (**Fig. 1c**), enhancing its accessibility for individuals from more deprived and historically underrepresented areas. Thus, current participants broadly reflect the sociodemographic characteristics of the UK adult population (see **Fig. 2**). A notable achievement has been the initiative's early success in ensuring that approximately 10% of participants come from UK ethnic-minority groups (excluding White minorities). Whilst this target means non-white participants are still underrepresented compared to the UK population, given its scale, the absolute impact of this representation is significant (**Fig. 2a**).

Although larger sample sizes can reduce random sampling error, a problem that all volunteer-based recruitment programmes face is that participant self-selection can nevertheless lead to selection biases. Transferability of genetic and other results therefore need to be carefully evaluated given that associations are modifiable and modifiers bias estimates towards those that are true for the over-represented group. The sample influences the association raising an external validity problem or, in other words, whether an association is observed in one study is dependent upon the distribution of the exposure-outcome relationship in the discovery and target population[3]. Notable examples include a 'healthy volunteer' bias in UKBB participants, which has been shown to distort genetic associations and downstream analys[4]. But some biases are more subtle, such as hidden geographic structure that has been shown to affect associations between genetics and complex traits[5].

Emerging weighting methods that make use of unbiased reference data are showing promise in correcting biases when modifiers are measured[1]. It is therefore encouraging that Our Future Health will be providing weights for the UK census 2021–2022 population and making these and the statistical code used to develop them available to researchers[2]. Yet, it is important to point

out that current weighting methods rely on phenotypic data, which may not correct participation bias on a genetic level and still lead to biased genotype-phenotype associations (see **Box 2**).

## Utilising the potential of genetic data and trans-biobank research

The available genetic data in Our Future Health (**Fig. 1b**) offers promising opportunities to improve our understanding of complex traits in exciting ways, particularly through future potential trans (or cross)-biobank research (i.e., combining data across biobanks)[9]. Although recruitment is ongoing, adding the current 650,979 Our Future Health genotyped participants to those of the UKBB (488K) and Genomics England's (GEL) sample (100K), for instance, the number already exceeds one million. If and when the Our Future Health sample reaches its target number of 5 million, genotyped individuals would reach approximately 10% of the UK adult population (**Supplementary Table 1**). As the fraction of the population that is genotyped increases, participants are increasingly likely to have at least one not-too-distant genotyped relative (i.e., those separated by 3 to 20 or more meiosis) in the dataset—this inflection point was demonstrated with Icelandic data fifteen years ago[10], which showed that when 10% of the population is covered, nearly all individuals had at least one such genotyped relative.

Having a high level of relatedness in a genetic dataset offers substantial analytical benefits, including the ability to separate out the paternally and maternally inherited haploid genomes for an individual. Although somewhat technical, this process of haplotype phasing is incredibly important to genetic research, as it, for instance, enables the improved study parent-of-origin effects[12]. Whilst information on haplotype origins is not available from the raw data, it can be deduced using methods that cross-match genetic data of not-too-distant related individuals. This was demonstrated for over 95% of Icelandic samples[10], and more recently, for over 120,000 UKBB samples[13]. Given the scale of Our Future Health, even without special effort in ascertaining first-degree relatives (i.e., sibling and parent-offspring pairs), which are increasingly recognised as important for many aspects of research[11], the cohort would still automatically comprise a large number of such relatives.

If Our Future Health, the UKBB, and GEL joined forces, the percentage of UK samples that could be successfully phased would increase when the number of genotyped individuals increases and their data are processed together. Large-scale datasets also tend to have less ascertainment bias, and importantly, the data themselves would provide more information to illuminate the nature of any bias and offer ways to make proper adjustments (**Box 2**). However, although such trans-biobank research would also offer many additional conceptual benefits[9,14], it comes with significant technical and regulatory challenges, all of which requires deeper investigation (see **Box 3**).

## Complementing national surveys and improving data linkage

The scale of Our Future Health presents another key opportunity: to complement existing population-level data efforts, such as the census, and enhance—or even partially replace—certain national surveys. A persistent challenge for these smaller surveys is declining response rates. In the UK, for example, the Office for National Statistics (ONS) has struggled with a drop in responses to the Labour Force Survey (https://www.resolutionfoundation.org/publications/measuring-up/), leading to increased

sampling variability. Similarly, the Health Survey England (https://digital.nhs.uk/data-and-information/publications/statistical/health-survey-for-england/2021/introduction) saw household response rates in 2021 fall by nearly 50% compared to 2019.

In contrast, Our Future Health collects detailed demographic, health, work, and lifestyle data at enrolment and plans to re-contact participants, offering a dynamic view of the population. While response rates may decline over time, it's recruitment model makes use of NHS (National Health Service) services (https://digital.nhs.uk/services/nhs-digitrials), offering unparalleled reach. The Netherlands and the Nordics have already embraced similar approaches, replacing traditional census data collection with population registries and biobank-linked datasets, integrated via unique personal identifiers. Though a UK census replacement is unlikely due to the absence of a national ID system and because linkage with administrative data (e.g., education, tax, employment) is governed under a separate legal framework to health data, such models offer lessons for Our Future Health in driving progress toward more integrated data. We acknowledge, however, that to achieve this, administrative data would need to be included in participant consent and other legal and technical challenges which also hamper trans-biobank remain (**Box 3**).

Notably, Our Future Health already links to electronic health records (EHRs) in the form of NHS records (see **Box 1**). As these can contain longitudinal geolocation indicators (i.e., participant location data), they could be used to link to other geolinked data on, for instance, institutional and policy measures on health or environmental indicators, such as wastewater pathogens or pollution, from local to national, across time and birth cohort. To safeguard participant privacy, data could be released at various levels of geographical area, such as Lower layer Super Outputs Areas (https://www.ons.gov.uk/methodology/geography/ukgeographies/statisticalgeographies) (LSOAs), which comprise between 400-1,200 households. This would allow Our Future Health to truly advance a precision public health approach, taking the environmental and social determinants of health into account[1].

To maximise the value of data linkage and cross-cohort analyses, the adoption of a common data model to ensure efficient mapping will be key. For illustration, comparing the prevalence of self-reported health conditions reported in the Our Future Health baseline questionnaire (which alone includes over 2,500 response codes) with national estimates, such as those from the Global Burden of Disease (GBD) Study (https://www.healthdata.org/research-analysis/gbd), is currently hindered by inconsistent mapping to GBD codes (**Fig. 3**). Ensuring interoperability between baseline questionnaire and clinical data with GBD, other existing international classification systems such as ICD-10 (https://icd.who.int/browse10/2019/en), ICD-O-3 (https://www.who.int/standards/classifications/other-classifications/international-classification-of-diseases-for-oncology), SNOMED CT (https://www.snomed.org/)), as well as those built on top (e.g., phecodes (https://phenomics.va.ornl.gov/phecodemap/)), will require additional curation. Collaboration with the UK research community (e.g., Health Data Research UK (https://www.hdruk.ac.uk/)) can help guide this process. One promising framework for this work is the OMOP Common Data Model[15], which has already been used by the UK Biobank and NHS initiatives to support replication, code reusability, and data integration in a federated way[14].

## Informing and actively learning from related global efforts

Whilst this Comment focuses on Our Future Health and the UK, which has proactively leveraged its population data to study and improve health, many points apply globally. The All of Us Research Program (https://allofus.nih.gov/), for instance, aims to enrol and collect a diverse group of at least one million individuals across the United States to accurately reflect the U.S. population. Global efforts are also growing, including the Qatar Biobank (https://www.qphi.org.qa/research/qphi-population-based-study) (QBB), the Abu Dhabi Biobank (https://abudhabibiobank.ae/), and those emerging in LMICs like GenomeIndia (https://www.nature.com/articles/d44151-025-00024-2), the Nigerian 100K Genome Project (https://www.nature.com/articles/s41588-022-01071-6), and Uganda Genome Resource (https://globalgenomics.org/ihcc-cohort-spotlight-uganda-genome-resource-ugr/) (UGR). Going forward, Our Future Health can help shape the global genomics landscape by engaging with international networks and consortia such as the Global Alliance for Genomics & Health (https://www.ga4gh.org/), the International Health Cohorts Consortium (https://globalgenomics.org/ihcc/), and the Global Biobank Meta-analysis Initiative (https://www.globalbiobankmeta.org/), alongside adopting the WHO's data sharing best practices (https://www.who.int/publications/i/item/9789240102149).

## Concluding remarks

To fully realise its scientific potential and shape the future of research, we support four priorities for Our Future Health from a researcher's standpoint. First, continue to pursue the 5 million target and recruit diverse communities, while also recognising bias through systematic investigation of phenotypic and genetic participation bias (**Box 2**), including clear documentation of any bias. Second, pursue broader data linkage (e.g., administrative data) but also enhanced documentation and data harmonisation, to enable cross-biobank comparability, complement national surveys, and enable future trans-biobank research**.** Third, address practical constraints of cloud genomics through open dialogue with researchers to assess the benefits and limitations of TREs, while ensuring privacy and security for participants (**Box 3**). Finally, prioritise proteomic and metabolomic data releases to allow researchers from across the life sciences to capitalise on this resource.

## Acknowledgements

VJS and MCM are supported by ESSGN (HORIZON-MSCA-DN-2021 (101073237), AK, SB and MCM by ESRC/UKRI Connecting Generations (ES/W002116/1). MCM by an ERC Advanced Grant (835079), EU MapIneq (202061645) and Einstein Foundation Berlin (EZ-2019-555-2). All authors by Leverhulme Trust Large Centre Grant LCDS (RC-2018-003). We thank the editor and reviewers for their constructive feedback.

## Competing interests

VJS is a Research Scholar for Our Future Health. MCM is a Research Ambassador for Our Future Health, a Trustee and on the Ethics Advisory Board of the UK Biobank, on the Scientific

and Ethics Advisory Boards of Our Future Health and Netherlands Lifelines Biobank, and on the Data Management Advisory Board of the US Health and Retirement Survey and UK CLS Cohort Studies. SB is a part of a working group called Alzheimer diagnostics that has received a grant from the Icelandic Technology Development fund. AK declares no competing interests.

## Boxes

**Box 1: Background on Our Future Health**

Below we provide a summary of Our Future Health (for details, see the Our Future Health research protocol (https://ourfuturehealth.org.uk/our-research-mission/how-our-future-health-works/)).

*What is Our Future Health?*

Our Future Health is an ongoing UK-based collaboration between the public, charity, and private sectors established in 2019 to build the UK's largest ever health research programme by recruiting up to 5 million participants, which began recruitment in October 2022 (**Fig. 1a**).

*What is its governance and funding model?*

Our Future Health is a charity governed by a Board of Trustees and supported by funding (https://ourfuturehealth.org.uk/about-us/how-we-are-funded/) from the UK government, industry, and medical charities. Initial 5-year base funding was provided by UK Research and Innovation (UKRI), a UK government-funding body, to set-up and deliver the programme. This was then supplemented by additional investment from the UKRI, the UK government, and multiple other sources, including additional funding from industry and charity partners.

*What is its recruitment strategy?*

Our Future Health recruits participants using community strategies, such as pharmacy collaborations, and partnerships with the UK's National Healthcare Service (NHS). Recruitment centres, 'pop-up' clinics, are located in shopping centres, mobile units, and supermarket car parks. Participants are recruited via post and email invitation letters sent to households residing near a clinic, alongside in-person advertisement at clinics. Every adult resident in the UK aged 18 or over can participate.

*What sets Our Future Health apart from existing efforts?*

Whilst it builds on the success of biobanks like the UK Biobank, what sets Our Future Health apart is both its absolute size, broader age range beyond older individuals, and the fraction of the population it will cover (**Supplementary Table 1**). Similar to Genomics England, Our Future Health is also embedded more closely within the NHS to ensure findings are translated into clinical practice. Additionally, it aims to make extensive use of a digital platform to keep contact with participants, recruit participants for further trials, and is developing a re-contact study framework to return personal disease risk information to participants.

*What information is collected and available to researchers?*

Our Future Health provides two main types of resources for researchers: (1) a prospective, observational cohort for basic scientific research, and (2), a translational research platform comprising a cohort of participants who can be re-contacted for follow-up. Available data (https://research.ourfuturehealth.org.uk/data-and-cohort/) at the time of writing includes questionnaire, genotype array, linked health records (i.e., secondary care including hospital stays, cancer registry, and deaths registry), and clinic measurements (e.g., blood pressure). Upcoming data to be released includes imputed genotype data, primary care linkage, and dispensed medications. Future data releases may include additional surveys and data linkages (e.g., disease registries), wearable data, whole-genome sequencing (WGS), alongside proteomic and metabolomic data.

*Who can access the data and resources?*

All researchers around the world who work for, or are a supervised student of, universities, charities, health services or organisations involved in health research can apply to access data by applying to become a 'registered researcher' and submitting a study application. Researchers from outside the UK and European Economic Area need to additionally provide an International Data Transfer Agreement (see Our Future Health territories of access (https://research.ourfuturehealth.org.uk/territories-of-access/)). Once successful, applicants are required to pay a fee and computing costs.

*What is the data sharing model?*

Our Future Health makes data and combined tooling available to registered researchers within its own Trusted Research Environment (TRE), a cloud-based data storage and analysis.

platform, as opposed to sending researchers extracts of data via a distributed sharing model. This TRE approach has become a standard of many biobanks.

**Box 2: Participation bias in genetic studies**

Participation bias refers to the phenomenon whereby a dataset is not representative of the intended study population. Such bias can distort prevalence estimates and affect downstream analyses in all types of sample surveys[3]. For genetic studies, samples are inherently biased given that participation requires consent to contributing DNA information. With genotype-phenotype associations, participation bias was until recently completely ignored for most studies, partly because of the hope that the effect would be negligible and partly due to the difficulties in quantifying and adjusting for the bias. That attitude, however, has begun to change and there is increasing awareness of the potential problems[6]. Various methods have been proposed to adjust results. Some methods require making strong assumptions about the nature of the participation bias, whilst others which rely on lesser assumptions would only be effective for large scale population-based studies.

One approach to adjust for participation bias utilises covariates that are available in both the genetic sample and census data (or a non-genetic sample that is presumably unbiased). Based on the sample-census difference, a propensity score is constructed to capture the relative

likelihood of participation and analyses are performed by attaching weights to samples that are inversely proportional to the propensity score[4]. One distinct advantage is that one set of weights serves all analyses. The flip side is that weighting samples increases standard errors, and using universal weights means that for one estimate, the price of statistical efficiency is paid for the proper adjustments of all other estimates. Moreover, if the propensity score is based on phenotypes only, it may inadequately adjust genotype-phenotype associations unless the genetic effects on participation is manifested entirely through the propensity score—a condition that was recently shown not to hold for one score constructed for UK Biobank (UKBB) data[7].

Genetic data have the property, possibly unique, that the sample alone contains information on the potential genetic basis of participation bias. In the UKBB, this was demonstrated by studying close relative pairs such as siblings and parent-offspring[8]. The existence of close relatives meant that shared segments participated twice (in two individuals) while non-shared segments only participated once. Alleles on shared IBD (identity by descent) segments between the relatives were found to be enriched for allele type that promotes participation compared to those alleles on segments not shared, indicating a heritable component to participation. Building on this, Song and colleagues[7] combined information on these genetic differences with phenotypic differences between the sample and a reference population to adjust estimates of heritability and genetic correlation for various phenotypes.

Although the extent of participation bias in Our Future Health remains to be fully characterised, its scale presents valuable opportunities. Participation is a complex trait and true allele frequency differences between shared and non-shared segments tend to be quite small for individual variants. Yet, adequate statistical power can be acquired by accumulating information from many studied variants (e.g., through the use of polygenic scores) and a dataset with a larger number of related individuals. Our Future Health may therefore provide an ideal platform to test and extend these methods, helping to correct any biases in the dataset and advancing methodological research on participation bias more broadly.

**Box 3: Benefits and challenges of trans-biobank research**

**Benefits**

***1. Estimating and resolving participant overlap***

Trans-biobank genetic analysis can estimate participant overlap between biobanks—a critical step for minimising bias in replication and meta-analyses and improving data integrity. Overlap checks could also uncover discrepancies in reported traits or health records for the same participants across biobanks such as Our Future Health and the UK Biobank (UKBB), creating an opportunity for harmonising data and highlighting errors.

**2. *Increased power and representation***

Cross-biobank analysis enhances statistical power, enabling better detection of rare variants, novel genotype–phenotype associations, and improved resolution for understudied populations for which sample sizes have been insufficiently small. For instance, whilst participants of South Asian ancestry represent approximately 24% of the world's population, until 2018 they made up only 2% of published GWAS (https://www.gwasdiversitymonitor.com/). Combining current genotype data on South Asian individuals in Our Future Health and the UKBB with data in Genes & Health (https://www.genesandhealth.org/), a smaller cohort of exclusively South Asian ancestry individuals (which currently holds data on >50K), would generate the world's largest South Asian ancestry cohort (>160K).

***3. Building long-term research infrastructure***

Allowing cross-biobank analysis builds the foundation for federated analysis pipelines and harmonised health studies. This is especially relevant in the UK, where analysing genetic data from Our Future Health with UKBB and Genomics England and beyond could in theory cover >10% of the adult population, enhancing population-scale genomic medicine. Although these biobank resources are highly distinct and complementary, there are models such as the UK Longitudinal Linkage Collaboration (UK LLC) (https://ukllc.ac.uk/), which pools survey data on over 325,000 participant records from more than 20 established longitudinal population studies in a TRE.

**Challenges**

***1. Data harmonisation and access***

Trans-biobank research requires reconciling differences in phenotypic definitions, variable codings, consent structures, biospecimen processing, omics assays and access policies. Whilst guidelines like the World Health Organisation's (https://www.who.int/publications/i/item/9789240102149) (WHO) recently published best practice recommendations on genomic data sharing provide emerging frameworks, aligning these across large-scale cohorts remains non-trivial. Population diversity and ancestry composition also varies widely, with the need to account for population stratification and environment.

***2. Legal frameworks, privacy, and governance***

Biobanks are often subject to very different legal frameworks (e.g., HIPPA in the US, GDPR in EU) and have highly varied consent models related to data usage. Ethical and data governance frameworks would need to cope with jurisdictional data sovereignty laws, nature of consent, cross-border processing limitations, and variation in ethics review processes to clarify how participant data might be reused in cross-cohort analyses. Given the varied providence of the data, questions of authorship, intellectual property, and data ownership may also arise. However, technical solutions of allowing data to remain in its own jurisdiction but linkage on a separate platform, such as those developed by ODISSEI (https://odissei-data.nl/facility/odissei-for-data-providers/) in The Netherlands, show promise.

### *3. Cloud constraints, interoperability, exploratory, and complex research*

Biobanks may also use distinct cloud platforms (e.g., TREs) that are not interoperable and scripts may not be portable, have different data formats, or varying cost structures. Higher or unpredictable costs may also hamper research and inadvertently discourage exploratory, trial-and-error experimentation, often essential for serendipitous discoveries. Trans-biobank research will also raise new types of statistical and methodological complexity and issues. Issues will arise such as harmonisation across covariates, strategies for handling missing data or diversity of sample (e.g., age, sex, ancestry, geography), and biases in one cohort (e.g., participation bias) can distort pooled results or require complex weighting.

Although further research is needed to address each of these challenges, it is encouraging to see the growth of initiatives like the Federated European Genome-Phenome Archive (https://ega-archive.org/about/projects-and-funders/federated-ega/) and for Our Future Health to jointly call for scaling up work on genomic data sharing with other biobanks[13].

## Figure legends

**Fig. 1:** Overview of recruitment in Our Future Health and related biobanks as of June 2025. **a**, Number of genotyped participants in Our Future Health (data release 11) and other mega-biobanks: MVP, Million Veteran Program (https://www.mvp.va.gov/pwa/); FinnGen (https://www.finngen.fi/en); UKBB, UK Biobank (https://www.ukbiobank.ac.uk/); All of Us (https://allofus.nih.gov/); EstBB, Estonian Biobank (https://genomics.ut.ee/en/content/estonian-biobank); MoBa, Norwegian Mother, Father and Child Cohort Study (https://www.fhi.no/moba-en) ; deCODE (https://www.decode.com/); BBJ, Biobank Japan (https://biobankjp.org/en/#gsc.tab=0); MCPS, Mexico City Prospective Study (https://www.ctsu.ox.ac.uk/research/prospective-blood-based-study-of-150-000-individuals-in-mexico); TWB, Taiwan Biobank (https://www.biobank.org.tw/english.php); CKB, China Kadoorie Biobank (https://www.ckbiobank.org/). **b**, Participants with baseline questionnaire data and projected future growth in Our Future Health, All of Us, and UK Biobank based on past recruitment rates. Shaded area represents potential future growth based on past enrolment rate. **c**, Recruitment centres of Our Future Health and UK Biobank, contextualised within regional life expectancy variations in Great Britain. *The number of open Our Future Health clinics corresponds to the number of clinic locations in England, Scotland, and Wales open at the time of writing. Data sources: Our Future Health Data releases (https://ourfuturehealth.gitbook.io/our-future-health/data/data-releases); All of Us Data snapshots (https://www.researchallofus.org/data-tools/data-snapshots/); Our Future Health interactive map (https://www.google.com/maps/d/u/0/viewer?mid=1zypNVdA0ckzOCxC9P-n9wJkVVeK6pwA&ll=51.346714600000034%2C1.0428765000000118&z=8).

**Fig. 2:** Self-reported sociodemographic characteristics of Our Future Health participants (data release 11) compared to the adult UK population using the 2021/2022 census and the UK Biobank. **a**, Ethnicity. Currently, over 95,000 participants identify as Asian British or any other Asian/Asian British background, over 26,000 as Black African, Black Caribbean, or other Black backgrounds, 4,000 as Arab, and over 30,000 as Mixed or multiple ethnic background. **b**, Sex (recorded as 'sex registered at birth' in Our Future Health; 0.1% of participants reported 'Other'). **c**, Highest level of educational qualification achieved in England and Wales and Our Future Health, coded by mapping the highest level of education that a participant reported to UK qualification levels (to ensure a more robust comparison, data is limited to participants aged 25-64 in both populations, and only includes Our Future Health participants born in

England or Wales; n=948,931). Percentages may not add up to 100% as participants reporting no qualifications are excluded. **d**, Age. Data sources: Office of National Statistics (https://www.ons.gov.uk/datasets/create/filter-outputs/c45dc5bb-9b6d-4be9-8065-380a5a4f9f78#get-data).

**Fig. 3:** Mapping the prevalence of common health conditions in the UK and Our Future Health (OFH) using participant self-report. Accurately harmonising OFH questionnaire data with external datasets will require standardised mappings. In this illustrative example, green circles represent mappings involving health conditions used by OFH that can currently be matched one-to-one (e.g., 'Cancer') to Level 3 causes from the Global Burden of Disease (GBD) Study. Orange circles represent mappings between health conditions in OFH (e.g., reproductive system problems) and GBD Level 3 causes in GBD (e.g., gynaecological disorders) that do not match one-to-one, highlighting the need for additional data curation and developing mappings. National prevalence of each health condition is estimated by the GBD Study 2021 for the United Kingdom. Using self-reported health conditions diagnosed by a medical professional, prevalence in OFH was estimated from n=1,781,893 participants aged 20 and above who have completed version 1 or 2 of the baseline health questionnaire. Data sources: GDB (https://www.healthdata.org/research-analysis/gbd).

## References


1. Gallagher, C. S., Ginsburg, G. S. & Musick, A. Biobanking with genetics shapes precision medicine and global health. *Nat. Rev. Genet.* **26**, 191–202 (2025).

2. Cook, M. B. et al. Our Future Health: a unique global resource for discovery and translational research. *Nat. Med*. **31**, 728–730 (2025).

3. Keyes, K. M. & Westreich, D. UK Biobank, big data, and the consequences of non-representativeness. *Lancet* **393**, 1297 (2019).

4. Schoeler, T. et al. Participation bias in the UK Biobank distorts genetic associations and downstream analyses. *Nat. Hum. Behav.* **7**, 1216–1227 (2023).

5. Abdellaoui, A., Dolan, C. V., Verweij, K. J. & Nivard, M. G. Gene–environment correlations across geographic regions affect genome-wide association studies. *Nat. Genet*. **54**, 1345–1354 (2022).

6. Munafò, M. R., Tilling, K., Taylor, A. E., Evans, D. M. & Davey Smith, G. Collider scope: when selection bias can substantially influence observed associations. *Int. J. Epidemiol.* **47**, 226–235 (2018).

7. Song, S., Benonisdottir, S., Liu, J.S. & Kong. A. Participation bias in the estimation of heritability and genetic correlation. *Proc. Natl. Acad. Sci. U.S.A.* **122** (25) e2425530122, https://doi.org/10.1073/pnas.2425530122 (2025).

8. Benonisdottir, S. & Kong, A. Studying the genetics of participation using footprints left on the ascertained genotypes. *Nat. Genet*. **55**, 1413–1420 (2023).

9. Deflaux, N., Selvaraj, M.S., Condon, H.R. et al. Demonstrating paths for unlocking the value of cloud genomics through cross cohort analysis. *Nat. Commun.* **14**, 5419 (2023).

10. Kong, A. et al. Detection of sharing by descent, long-range phasing and haplotype imputation. *Nat. Genet.* **40**, 1068–1075 (2008).

11. Davies, N. M. et al. The importance of family-based sampling for biobanks. *Nature* **634**, 795–803 (2024).

12. Kong, A. et al. Parental origin of sequence variants associated with complex diseases. *Nature* **462**, 868–874 (2009).

13. Kutalik, Z, et al. Parent-of-Origin inference and its role in the genetic architecture of complex traits: evidence from~ 265,000 individuals. (2025). Preprint at: https://www.researchsquare.com/article/rs-5871891/v1

14. Stark, Z. et al. A call to action to scale up research and clinical genomic data sharing. *Nat. Rev. Genet.* **26**, 141–147 (2025).
15. Reich, C. et al. OHDSI Standardized Vocabularies—a large-scale centralized reference ontology for international data harmonization. *J. Am. Med. Inform. Assoc.* **31**, 583–590 (2024).