biobank uk

Enabling scientific discoveries that improve human health

UK Biobank: democratising access to large-scale genomic and phenotypic data for discovery science

Dr Ben Lacey

UK Biobank Epidemiology Group

University of Oxford

N8 CIR: Data Access for Digital Health

8 Nov 2023

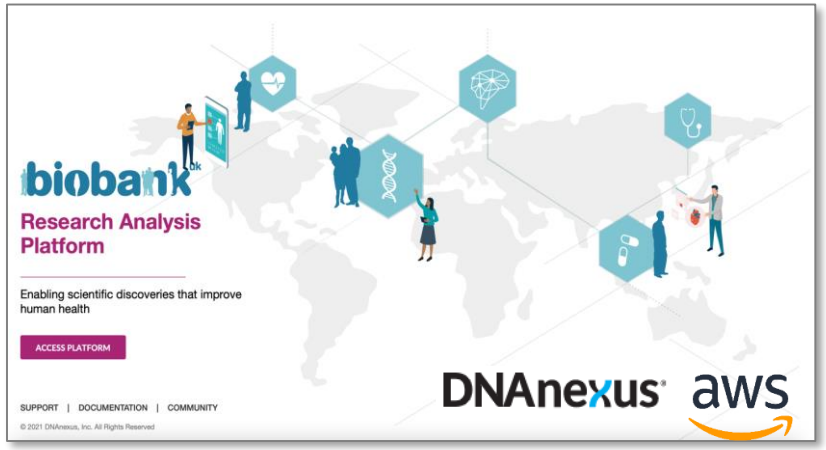
www.ukbiobank.ac.uk

What is UK Biobank?

- Prospective cohort study, following the health of 500,000 people to understand the genetic, physiological, lifestyle, and environmental determinants of disease

- UK Biobank has a **unique** combination of scale, depth, and duration of follow-up

- Research database to enable scientific discoveries, which is **readily accessible** to researchers worldwide



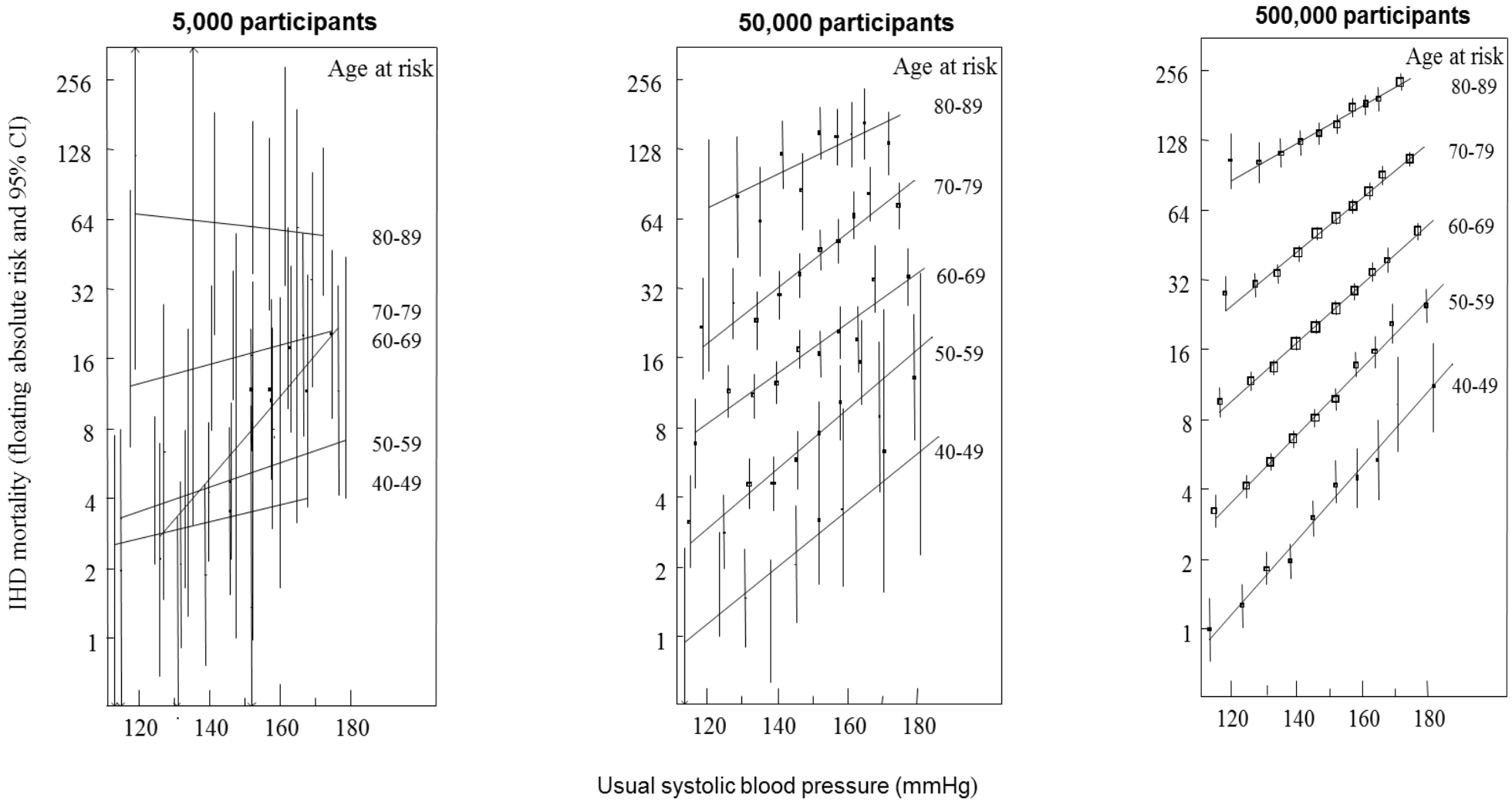
Why was UK Biobank established?

- The causes of most disease are due to a combination of many factors…genetics, physiological factors, lifestyle, environment (**detail** – we need a lot of data on participants)

- Moderate or small effects from common exposures, or large effects from rare exposures, are difficult to detect (**scale** - we need a lot of participants)

Need for a large-scale prospective cohort with deep characterization of participants and long follow-up

Why was UK Biobank established?



Data from the Prospective Studies Collaboration, 2002

Overview of UK Biobank recruitment

- Between 2006-2010
- Aged 40-69 years old
- Registered with the NHS
- Living within ~25 miles of 1 of the 22 assessment centres

Data collected at UK Biobank Recruitment

Touchscreen questionnaire	Demographics, environment, lifestyle, medical history, and cognitive function and hearing tests
Verbal interview	Occupation, medical conditions, medications, operations, etc.
Physical measures	Blood pressure, heart rate, anthropometry, spirometry etc.; and arterial stiffness, bone density, eye measures, fitness test (inc. ECG) in subset.
Sample collection	

Data collected at UK Biobank Recruitment

- Blood
- Urine
- Saliva

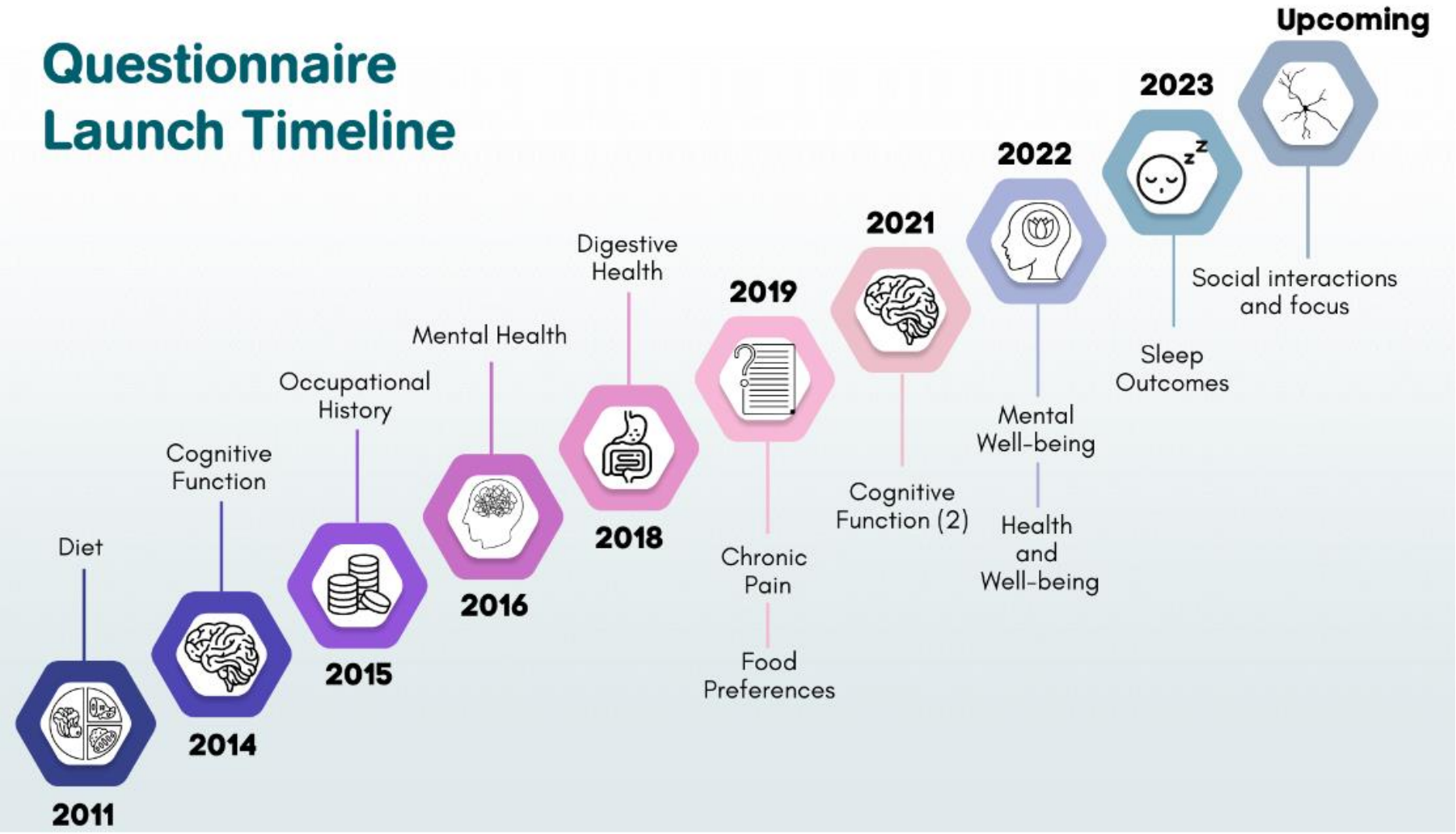
Total > 15 million aliquots

UK Biobank baseline characteristics

Characteristic	Category	Participants, n (%)
Age	40-49 yrs	119,000 (24%)
	50-59 yrs	168,000 (34%)
	60-69 yrs	213,000 (42%)
Sex	Male	230,000 (46%)
	Female	270,000 (54%)
Ethnicity	White	473,000 (95%)
	Other	27,000 (5%)
Deprivation	More	92,000 (18%)
	Average	166,000 (33%)
	Less	241,000 (46%)
Total		500,000

Wide range of backgrounds represented

Enhancements: Web-based questionnaires



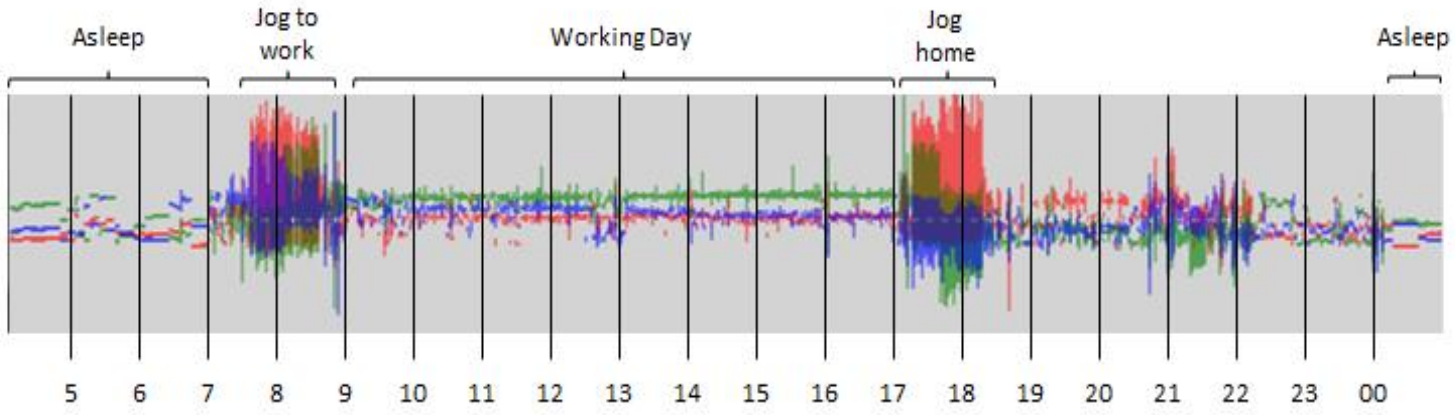
Among 330,000 participants for whom we have an email address.

Enhanced information on selected exposures and outcomes that is was not feasible to collect at baseline.

Enhancements: Physical activity monitor

Accelerometer data

- 100,000 participants
- Worn continuously for 7 days
- 2,500 people repeated seasonally



Enhancements: Resurvey and imaging

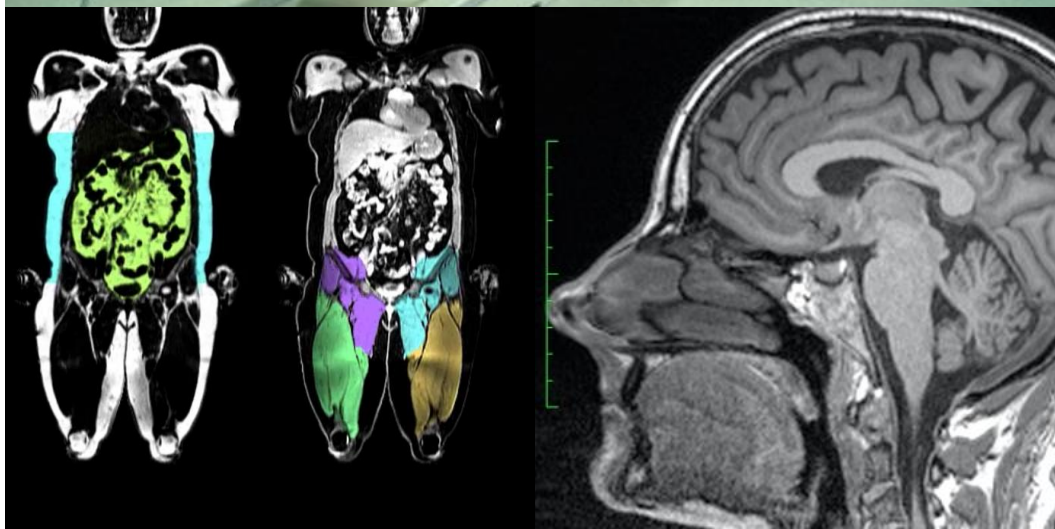
Repeat assessment in 20,000 participants (2012-13)

Multi-modal imaging (60,000 of 100,000 ppts; 2014-)

- MRI (heart, brain, abdomen)
- Full-body DEXA
- Carotid ultrasound
- 12-lead ECG

Repeat imaging underway (target of 60,000 ppts; 2019-)

The world's largest multi-modal imaging study



Enhancements: Samples into data

Genetics

Genome-wide genotyping

- 850k variants directly measured; >90M variants imputed
- 500,000 made available 2017

Whole Exome Sequencing

- 450,000 in Q4 2021; further 20,000 made available 2022

Whole Genome Sequencing

- First 200,000 made available Q4 2021
- Full cohort to be made available later in 2023

Enhancements: Samples into data

Biochemical measures in all 500,000

- 34 biomarkers in plasma, serum, red blood cells, and urine samples

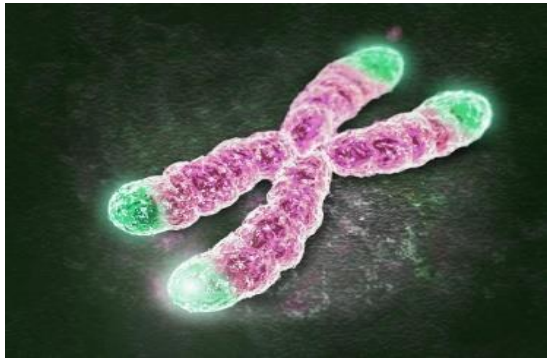
Telomere length in all 500,000

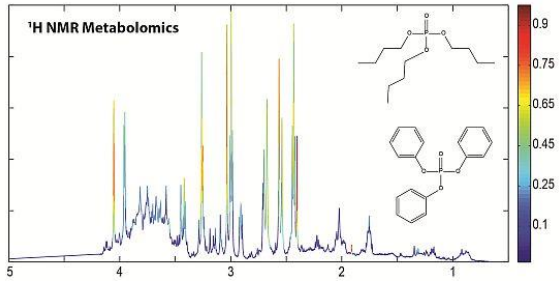
NMR-metabolomics in 300,000

- Data released for first 120,000 in 2021
- Second phase released Aug 2023

Proteomics in ~60,000

- Pharma consortium
- ~1500 plasma proteins using Olink's assay released
- Second phase release of further ~1500 plasma proteins later this year



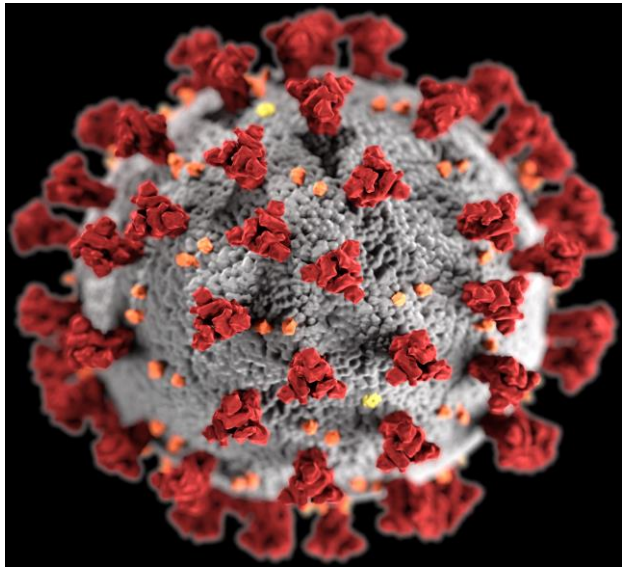


Enhancements: Enabling COVID-19 research

Continuing data collection and high participant engagement enabled mobilisation of resources for COVID-19 research

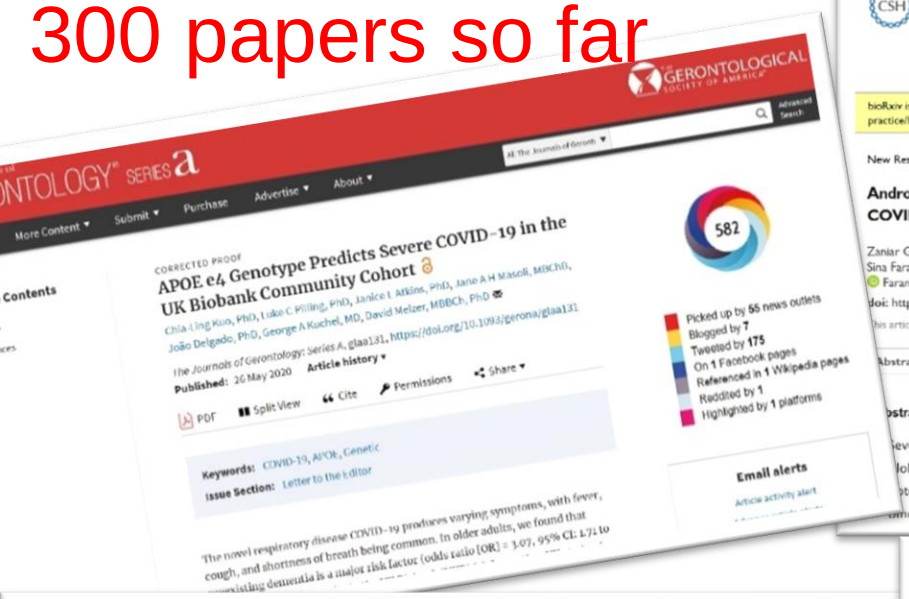
- Enhanced COVID-related health outcome data

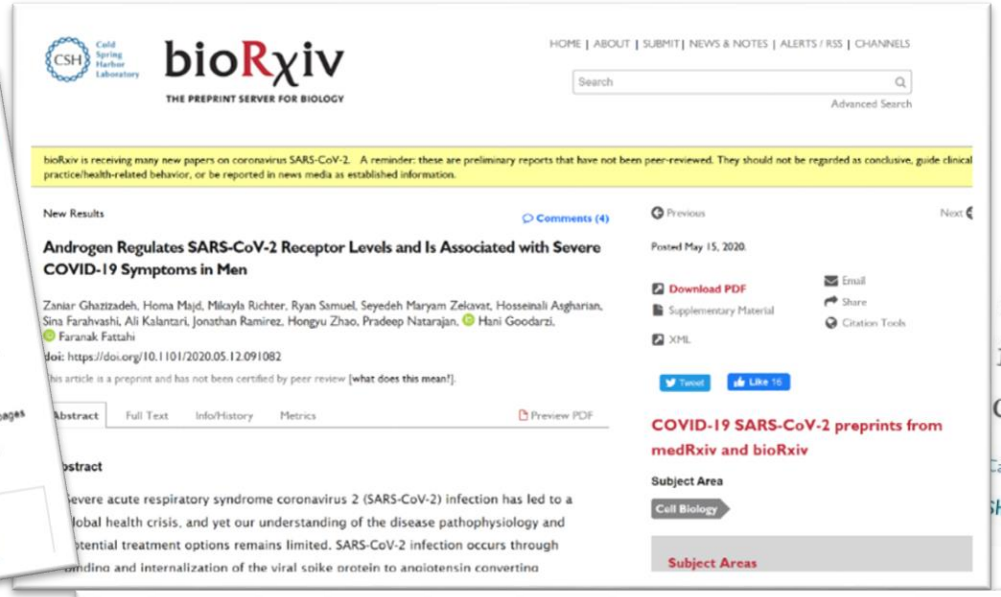
- Several COVID-related sub-studies
  - -UK Biobank COVID-19 serology study
  - -SARS-CoV-2 imaging study



UK Biobank researchers and publications on COVID-19

700 research groups
300 papers so far

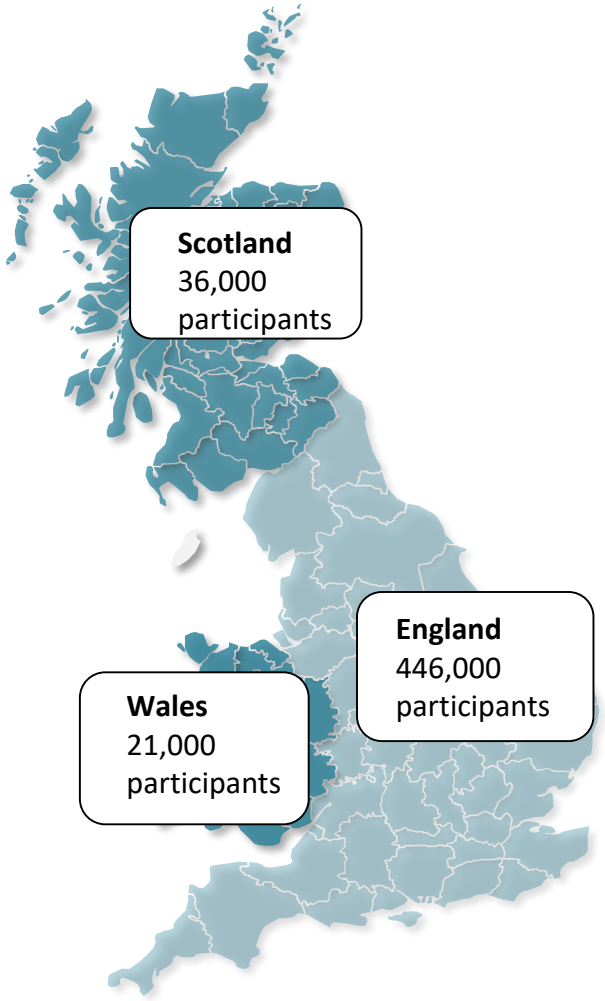




Follow-up of health outcomes

Regularly updated information on a wide range of diseases from NHS datasets in all 3 countries:

- **Deaths** (date and cause)
- **Cancers** (date and type)
- **Hospitalisations** (dates, diagnoses, procedures)
- **Primary care** (~45% of participants; dates, diagnoses)
- **COVID-related** (SARS-CoV-2 antigen tests)



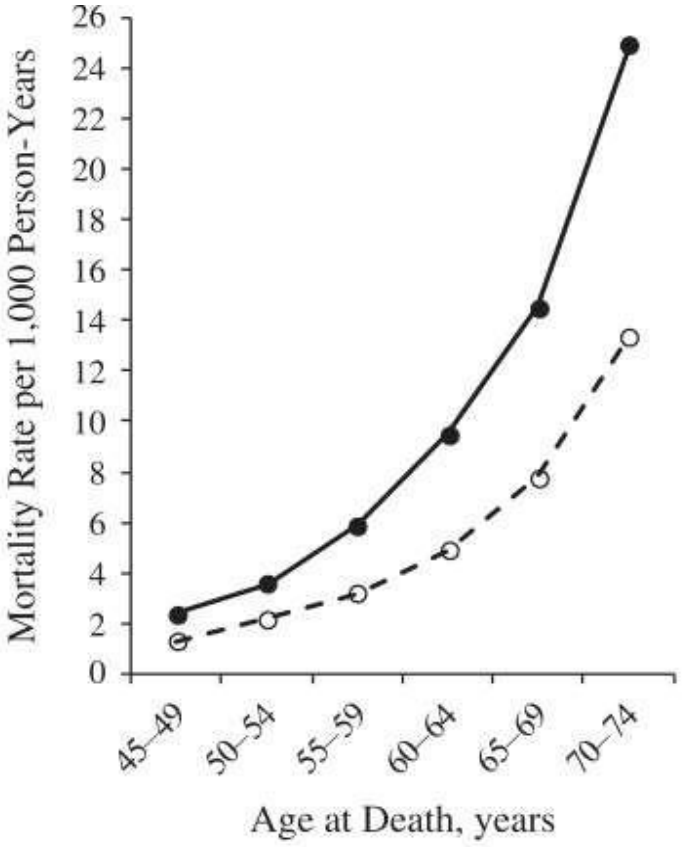
Cumulative number of incident case over time

Condition	Year of diagnosis		
	Observed	Predicted	
	2022	2027	2032
Diabetes	31,000	54,000	70,000
Myocardial infarction	15,000	30,000	46,000
Stroke	12,000	25,000	37,000
COPD	25,000	47,000	65,000
Depression	25,000	39,000	47,000
Breast cancer	9,000	14,000	18,000
Colorectal cancer	5,000	8,000	11,000
Lung cancer	4,000	6,000	8,000
Prostate cancer	10,000	16,000	20,000
Hip fracture	5,000	13,000	22,000
Rheumatoid arthritis	4,000	6,000	8,000
Parkinson’s disease	4,000	10,000	14,000
Alzheimer’s disease	5,000	17,000	37,000

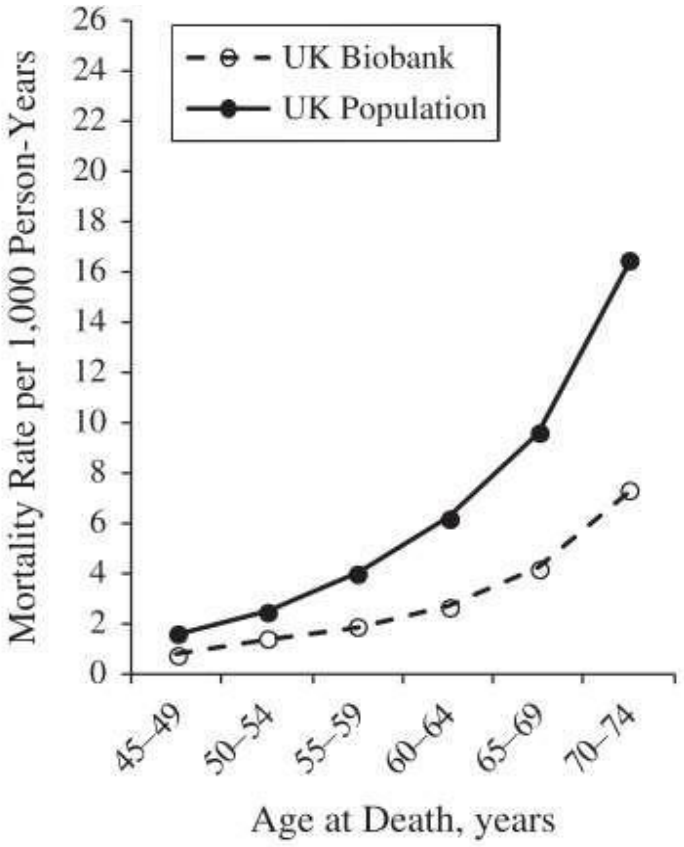
Mortality rates compared to the general population

Mortality rates ~50% lower in UK Biobank

Men



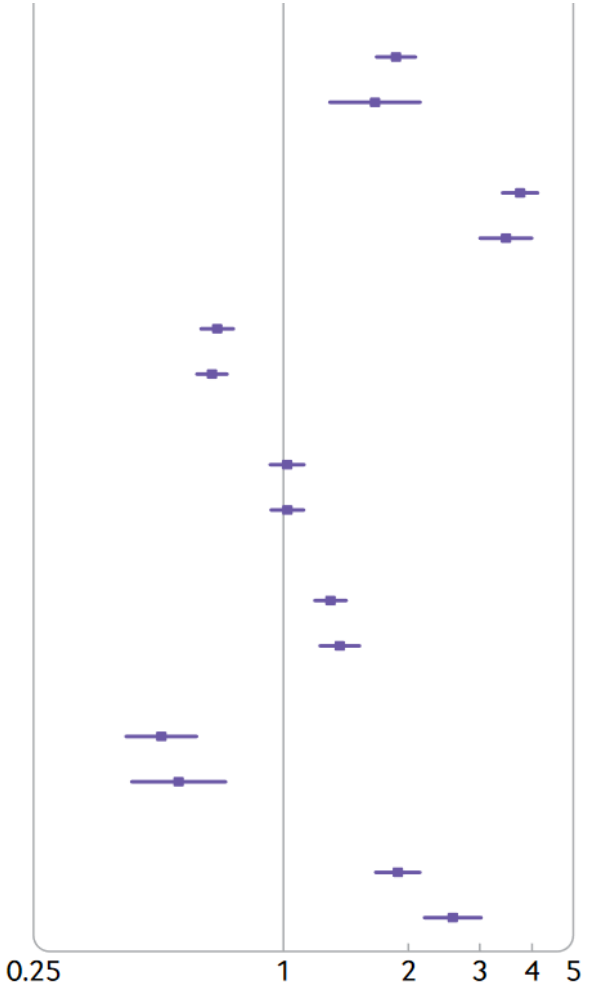
Women



Fry A et al. Am J Epidemiol. 2017;186(9):1026-1034

Relative risk in UK Biobank vs representative studies

Risk factor and study	No	Deaths	Hazard ratio (95% CI)
Elevated waist:hip ratio			
UK Biobank	502 655	2251	1.88 (1.69 to 2.09)
HSE-SHS	56 513	875	1.67 (1.31 to 2.14)
Self-reported diabetes			
UK Biobank	500 038	2233	3.73 (3.38 to 4.12)
HSE-SHS	89 871	1596	3.43 (2.96 to 3.96)
FEV_1 (per 0.5 L/s increase)			
UK Biobank	453 836	1765	0.69 (0.67 to 0.72)
HSE-SHS	26 742	629	0.68 (0.64 to 0.72)
Total cholesterol (per 1 mmol/L increase)			
UK Biobank	387 923	1074	1.01 (0.96 to 1.07)
HSE-SHS	30 946	477	1.04 (0.96 to 1.12)
C reactive protein (per doubling in mg/L)			
UK Biobank	449 141	1859	1.30 (1.25 to 1.35)
HSE-SHS	19 475	220	1.36 (1.23 to 1.52)
HDL cholesterol (per doubling in mmol/L)			
UK Biobank	354 782	995	0.51 (0.42 to 0.62)
HSE-SHS	25 951	323	0.56 (0.43 to 0.72)
Hypertension			
UK Biobank	494 047	2179	1.89 (1.69 to 2.11)
HSE-SHS	57 116	1002	2.56 (2.20 to 2.98)



Hazard ratios for cardiovascular mortality.

HSE, Health Surveys for England, SHS, Scottish Health Surveys.

Batty et al. BMJ 2020;368:m131

Enhancing health outcomes

Rationale for this work

- Important to identify as many cases as possible to increase the power of the study to investigate associations between risk factors and disease
- Reduce the number of false positive for disease, also important for study power
- Increase the range of diseases that can be studied
- To increase biological specificity of disease classification - increase the number of cases across the disease severity spectrum and across disease sub-types

Enhancing health outcomes

Expanding health linkages with routine healthcare data

Potential new linkages:

- Mental health services
- Psychological therapy
- Microbiology
- Joint register
- Microbiology
- Clinical disease audits
- Ophthalmology datasets
- Govt. data (income, education..)

Based on:

- scientific added-value
- data quality
- comprehensiveness
- cost-effectiveness
- feasibility

Enhancing health outcomes

Use of online questionnaires to obtain self-reported data on health outcomes

Future web-based questionnaires: in development

- Neurodevelopment disorders (ADHD, autistic traits)
- Cognitive function outcomes
- Neurological disorders (lack of facial recognition and mental imagery)

Enhancing health outcomes

Future efforts to identify dementia sub-types

- Potential to invite participants with diagnosis of dementia to an assessment clinic for:

- **Blood-based neurodegenerative biomarkers**
  - E.g. Ptau-181 and -271, (Aβ)42, neurofilament light change protein (NfL)

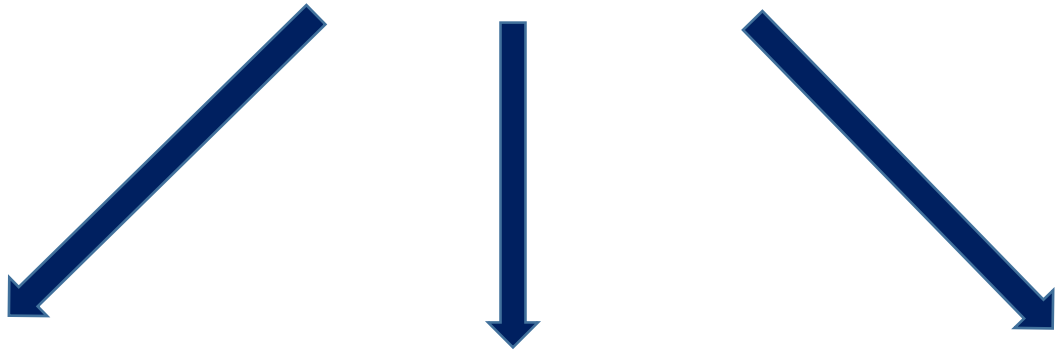
- **Imaging scans**

- **Clinical presentation**
  - Including possible use of apps

Enhancing health outcomes

Generation of ‘research-ready’ health outcomes

Diagnostic, procedural and health-related codes contained in diverse health care records and online questionnaires



Mapping coded data across a diverse range of health care records into a common standard (4-digit ICD-10; OMOP)

Integrate open-source tools and algorithms that define health outcomes (e.g., HDRUK Phenotype Library)

Support data challenges to explore AI-based approaches to classify health outcomes

Accessing the data

Principles of accessing UK Biobank

- Open access resource, available for bona fide researchers to conduct health-related research that is in the public interest
- Available for use by academia and industry, both in the UK and overseas
- No preferential or exclusive access to the resource (*and limited exclusive access for data generated by researchers*)
- Researchers are obliged to return their results to UK Biobank so they can be shared with others

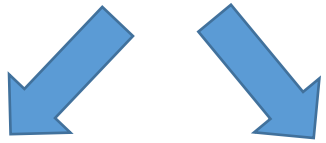
Accessing the data

Access procedure and options

Register and submit application

Application approved

Fees based on ‘tier’ of data requested
Tier 1: Clinic and follow-up data (£3k excl. VAT)
Tier 2: Above plus genotyping and other assay data (£6k excl. VAT)
Tier 3: Above plus sequencing and imaging data (£9k excl. VAT)



Download dataset to local computer

Access data in the Research Analysis Platform

- WES and WGS data only available here

UKB Research Analysis Platform (RAP)

>390
Community Members

>2,700
Projects

200+
Tools

>500
Active Users

3,297
Users

26+
Webinars and Workshops

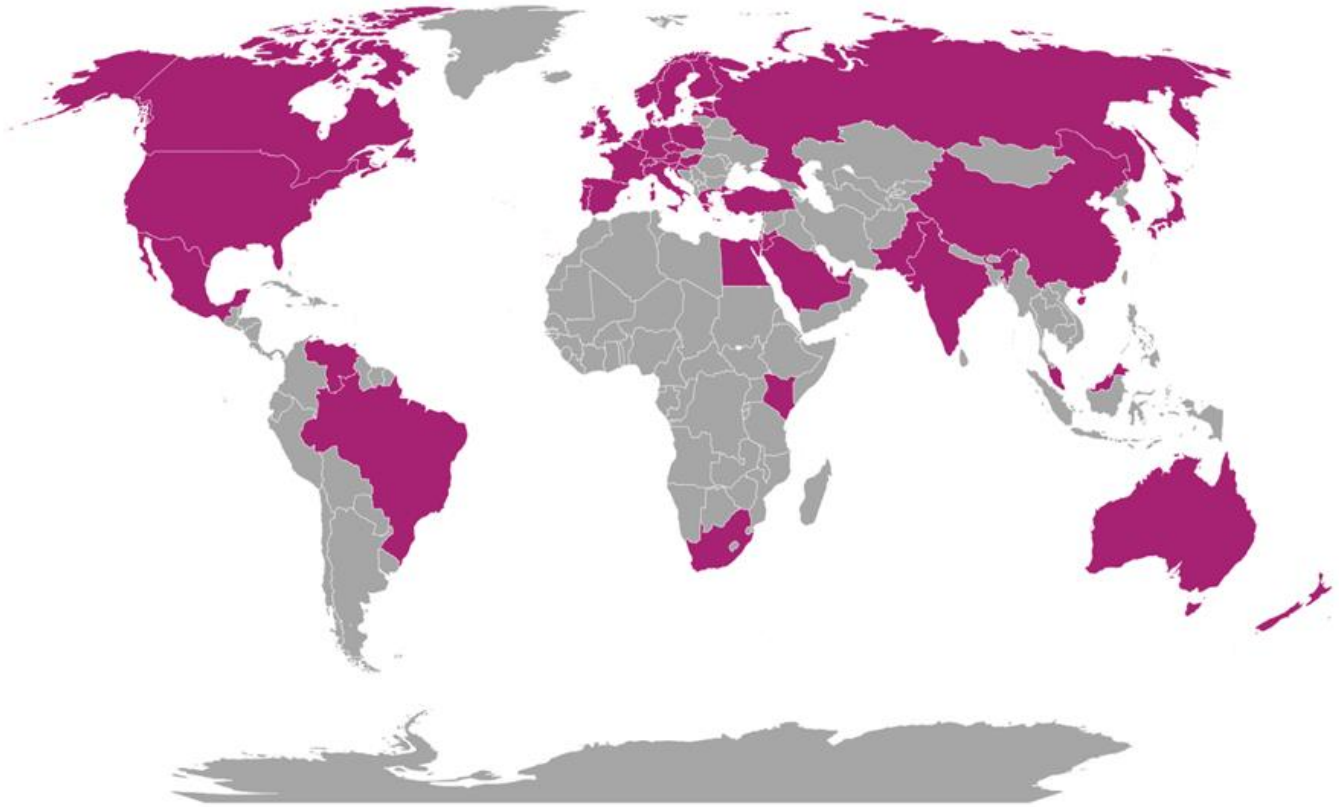
27.1PB
Data

11M
Files

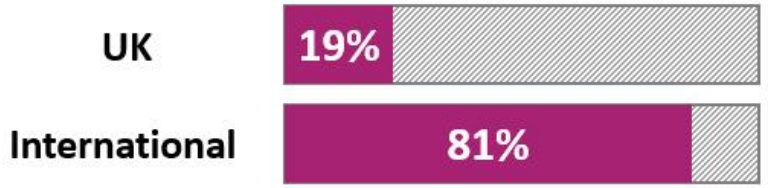
463+
Apps

Research credits available to LMIC and early career researchers

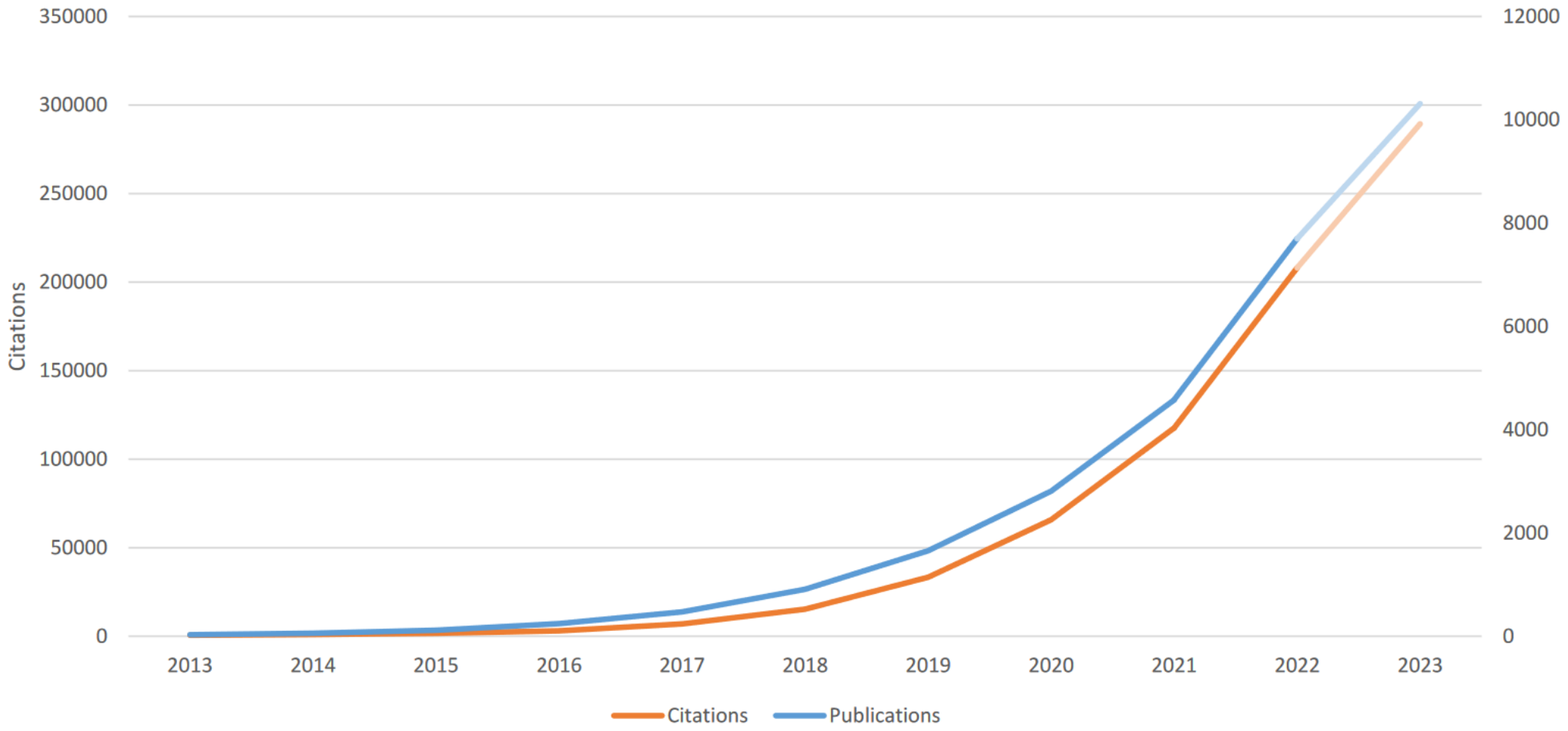
Who is using the UK Biobank dataset?



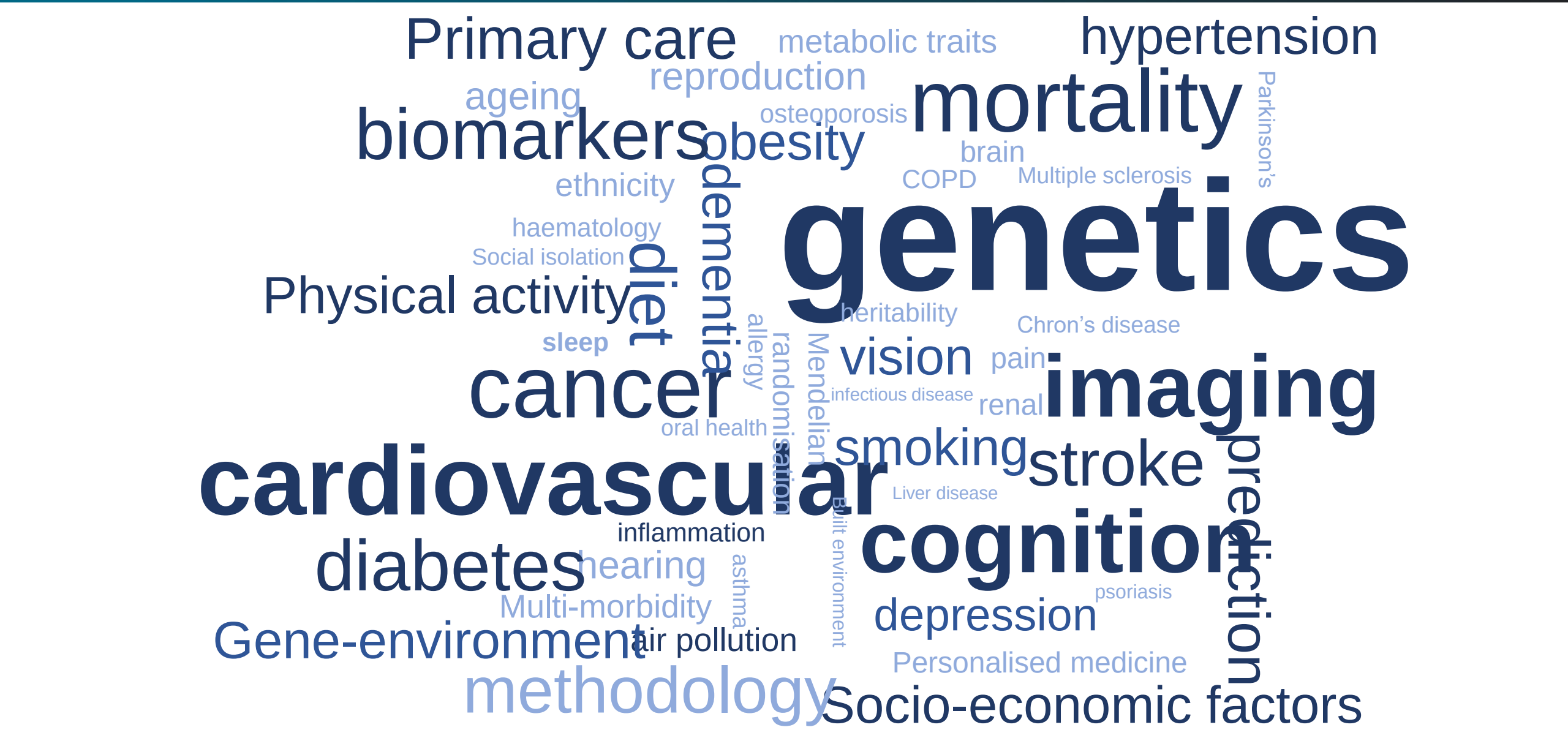
33,500 researchers
90 countries



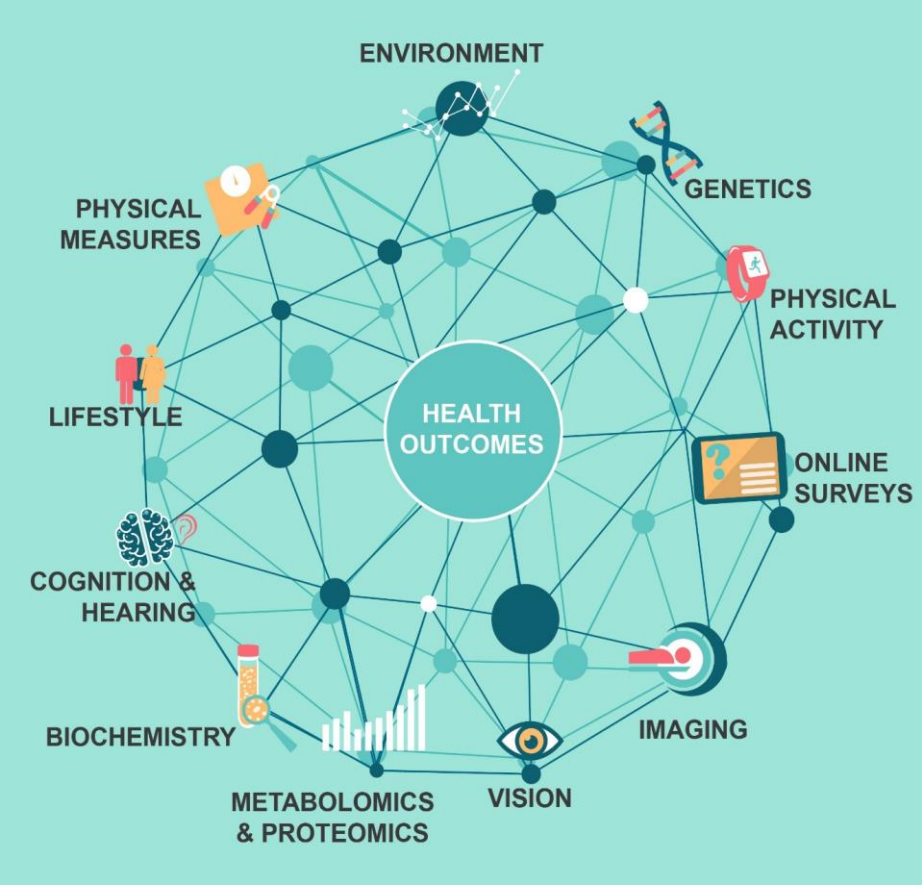
Publications and citations using UK Biobank



What research is being done?



UK Biobank: Unique combination of 4 dimensions



Combination of size with increasing depth x duration x accessibility enabling cutting-edge science

- SIZE: 500,000 diverse individuals
- DEPTH: Genetics with extensive detail about lifestyle, environment and medical history, and other biological assays (biochemistry, genetics, -omics) and imaging
- DURATION: >10 years of follow-up has already yielded very large numbers of many different health outcomes
- ACCESSIBILTY: Very rapidly increasing number of different types of researcher globally using UK Biobank, facilitated by the UKB RAP

Acknowledgments

Core Funders

UKRI Medical Research Council

wellcome

CANCER RESEARCH UK

British Heart Foundation

Department of Health & Social Care

UK Biobank Team and Collaborators

Executive Team and Coordinating Centre staff, Steering Committee, International Scientific Advisory Board, Scientific Working Groups, Oxford University Team

Our 500k participants