Obtaining phenotype and outcome data from e-health records and digital platforms: the experience of UK Biobank

Cathie Sudlow

Professor of Neurology and Clinical Epidemiology

Director, Centre for Medical Informatics, Usher Institute, University of Edinburgh

Director of Health Data Research UK Scottish substantive site

Chief Scientist, UK Biobank

International Cohorts Summit,

Durham, North Carolina

March 2018

UK Biobank in a nutshell

- 500,000 UK men and women aged 40-69 years when recruited during 2006-2010
- Consent for all types of health research by both academic and commercial researchers
- Extensive baseline questions and physical measures, with biological samples stored for future assays
- Subsequent enhancements in all or large subsets of participants:
  - Data from portable wearable devices (100,000 accelerometry; 20,000+ continuous ECG)
  - Sample assays in all or large subsets:

    Complete: genome-wide genotyping; biochemistry panel

    Underway/planned: exome and whole genome sequencing; proteomics; infectious disease assays; stool microbiome
  - Multimodal imaging of 100,000 (>22,000 so far)
  - Web questionnaires
- Comprehensive, long term follow-up for a wide range of health-related outcomes
- Open access for approved research: see www.ukbiobank.ac.uk

biobank uk

Follow-up of participants in very large prospective cohorts

Aim: identify a wide range of incident diseases and other health related outcomes

Active methods requiring participant re-engagement

- face to face reassessment
- postal or web-based surveys
- expensive
- prone to incomplete coverage & selective loss to follow-up
- miss cases emerging between assessments

Passive methods via linkages to national health records

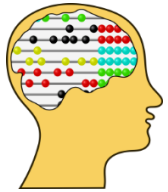
- can follow all participants without need for re-engagement
- efficient and cost effective
- need adequate consent at recruitment
- rely on universal healthcare system & availability of relevant datasets
- can only detect cases of disease diagnosed in a healthcare setting
- data need to be accurate and sufficiently detailed for research studies

biobank uk

Web questionnaires

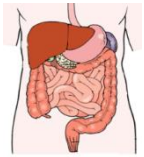
- Using email and web questionnaires
  - for more detailed assessment of exposures
  - and to obtain information on outcomes that cannot be obtained through linking to health records
- Of 350,000 with email, >150,000 complete each questionnaire

  - Details of dietary intake



  - Cognitive function

  - Mental health (thoughts and feelings)



  - Gastrointestinal symptoms

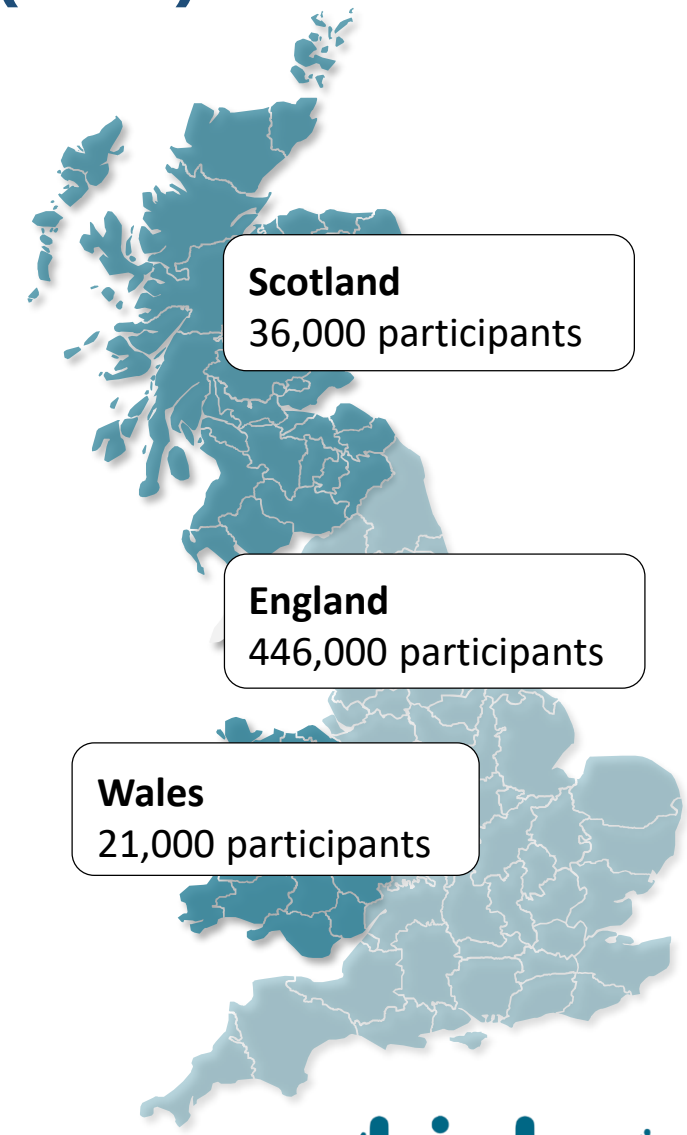
Useful for following change over time…but beware selective attrition

biobank uk

Following the health of 0.5 million UK Biobank participants through linking to National Health Service (NHS) records

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

- Deaths - date and cause of death
  for all participants
  >14,000 by early 2016
- Cancers – date, stage and grade of cancer
  for all participants
  >79,000 cancer cases by late 2015
- Admissions to hospital – dates, diagnoses, procedures
  for all participants
  1000's of cases of many incident diseases
- Primary care data – dates, diagnoses, symptoms, signs, referrals, prescriptions, labs etc
  for half of the participants
  1000's more cases of many incident diseases



biobank uk

Maximising the value of the linked healthcare data

- Messy ‘real world data’ - not collected primarily for research
- Not 100% accurate due to administrative and clinical error
- Mainly structured, coded datasets (ICD, OPCS4, Read…)
- Experts advising in a range of disease areas:

Cancer
Diabetes
Cardiac diseases
Stroke
Mental health disorders
Eye diseases
Neurodegenerative diseases
Chest diseases
Musculoskeletal conditions
Infections
Kidney diseases

- Combine different linked data sources to create algorithmically derived disease status indicators
- Estimate the accuracy and completeness of these
- Consider limitations and potential additional sources of unstructured data

biobank uk

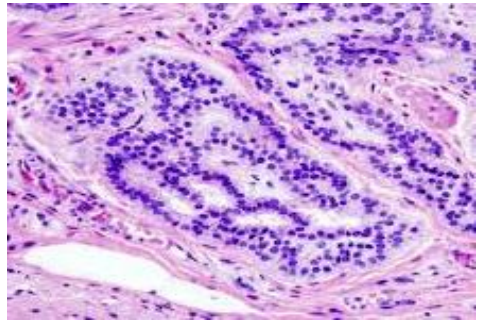
Cancers in UK Biobank ascertained from the national cancer registries

	Observed		Predicted
	By recruitment	Incident by 2015	Incident by 2022
Breast cancer	9,000	4,200	10,000
Colorectal cancer	2,300	2,500	7,000
Prostate cancer	3,000	4,300	9,000

→ Date, stage and grade of cancer

Beyond the structured registry data…exploring feasibility of retrieving additional information for subtyping of identified cancer cases through regional linkages to:

- histopathology reports
- digitised histopathology slides
- tumour specimens



biobank uk

Exemplar non-cancer conditions in UK Biobank

ascertained from baseline self report, hospital admissions and death registries

	Observed	
	By recruitment	**Incident by 2016**
Myocardial infarction	12,000	7,400
Stroke	8,000	4,600
Diabetes	26,000	9,000
COPD	10,000	7,600
Asthma	60,000	5,700
Dementia	200	1,800

biobank uk

Exemplar non-cancer conditions in UK Biobank

ascertained from baseline self report, hospital admissions and death registries

Estimated effects of including primary care data

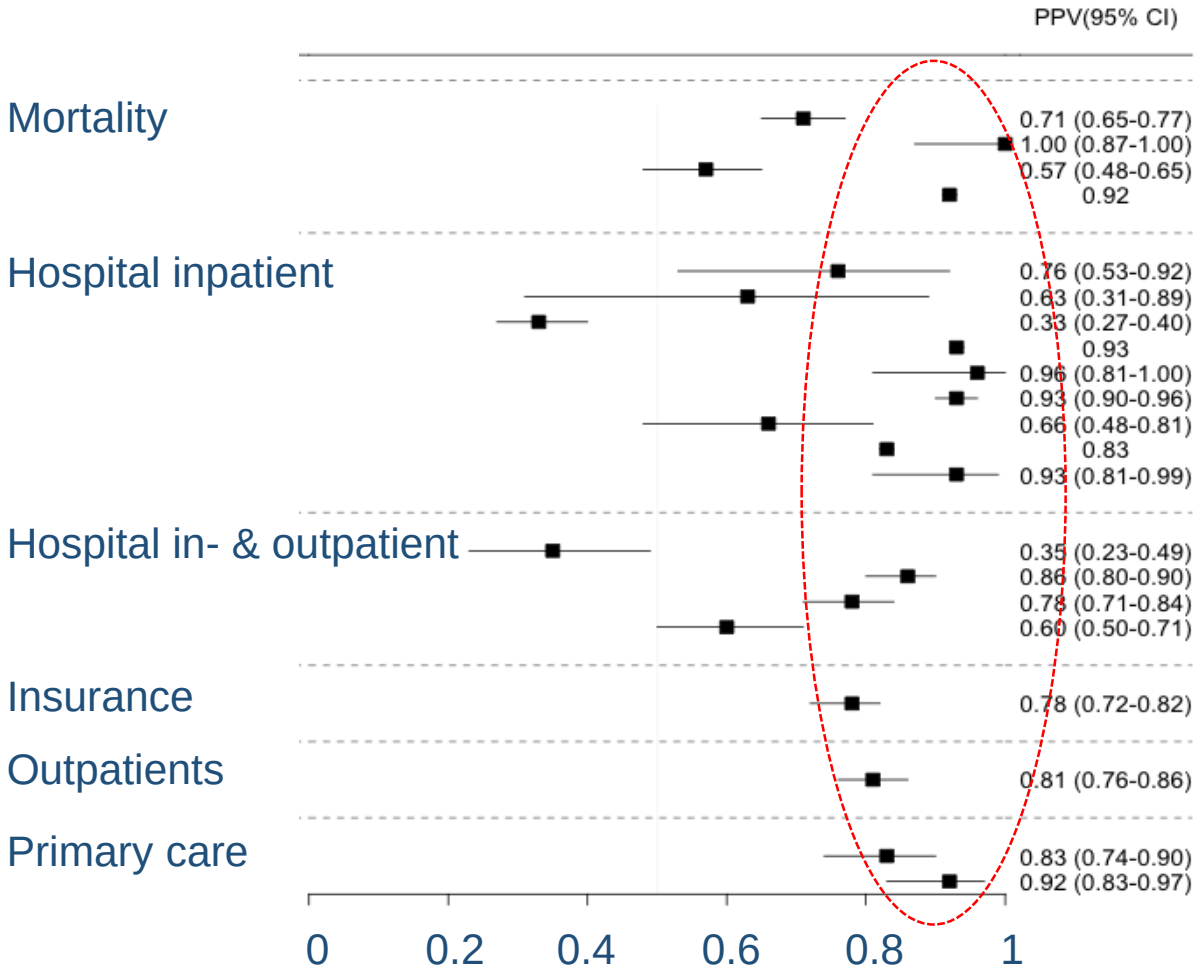
	Observed		
	By recruitment	Incident by 2016	
Myocardial infarction	12,000	7,400	8,100
Stroke	8,000	4,600	6,900
Diabetes	26,000	9,000	18,000
COPD	10,000	7,600	16,900
Asthma	60,000	5,700	19,000
Dementia	200	1,800	3,600

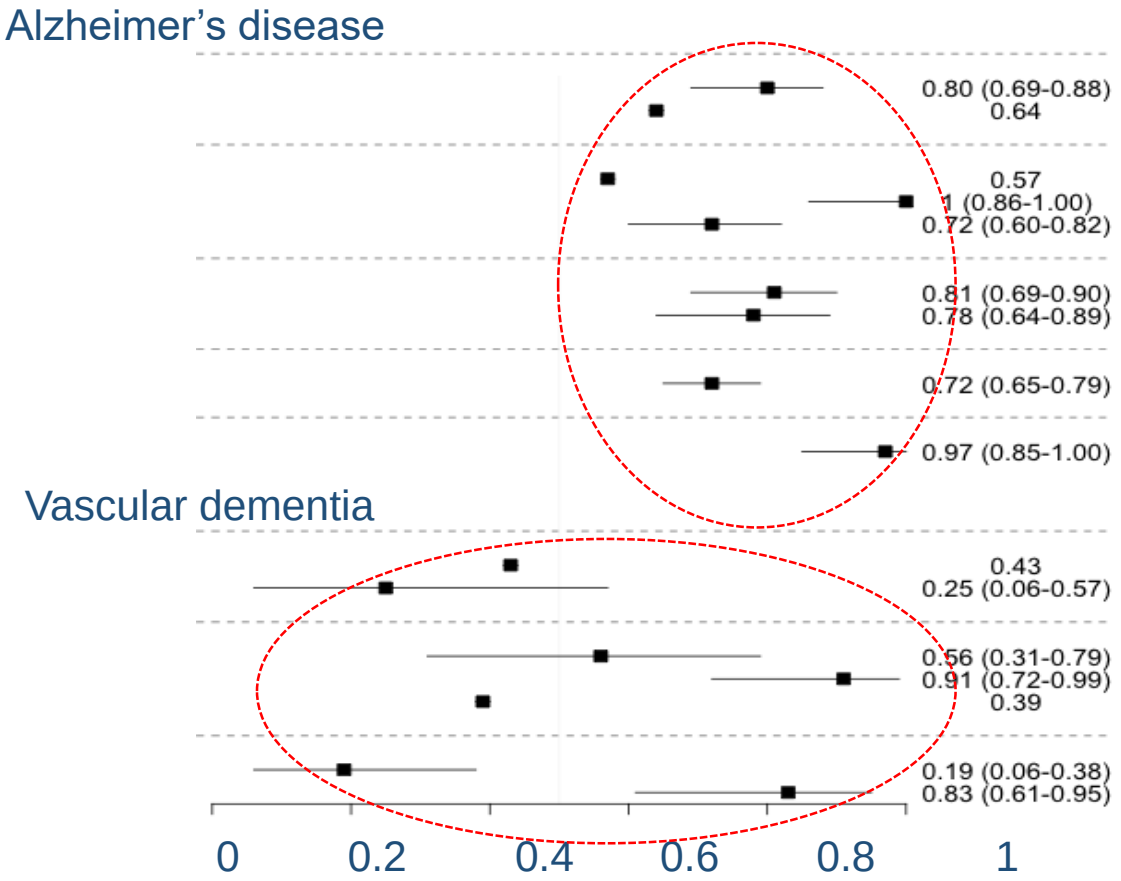
Accuracy?
Limitations?

biobank uk

Dementia: positive predictive value of routine healthcare data

From published studies





Wide variation but in most PPV >80%

PPV for AD generally higher than for vasc dementia

Dementia: positive predictive value of routine healthcare data

From comparison with expert review of free text electronic medical record in UK Biobank

Correspondence

Number of correspondence letters: 6

Letter 1

Setting:
Psychiatry letter, April 2015

I was pleased to assess Mr [redacted] during a liaison hospital visit to ward 3, Royal Victoria Hospital on [redacted].4.15. In addition to meeting Mr [redacted], I read his current multidisciplinary team notes, his past psychiatric notes and had the opportunity to discuss nursing management with Charge Nurse, [redacted].

Mr [redacted] has been in contact with my Community Psychiatric Nurse colleagues from my sector Community Mental Health Team for some time and I had received a recent request to offer medical review, although he had subsequently been admitted to the Western General Hospital. I have had discussions with [redacted] (Bridging Team Nurse) and a number of conversations with Mr [redacted]'s wife, who had phoned me directly. Mrs [redacted] had been concerned about his condition and she appeared to have been having difficulty coming to terms with the extent of his problems and was uncomfortable with the arrangements being discussed for future placement in either a nursing home or in-patient complex care hospital environment. She had, of course, been caring for

Dementia:	80%
Alzheimer's disease:	72%
Vascular dementia:	44%

biobank uk

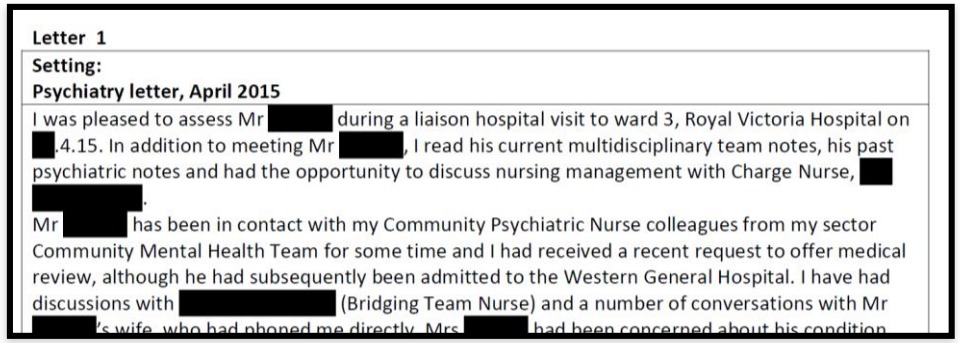
Beyond the linked coded healthcare data

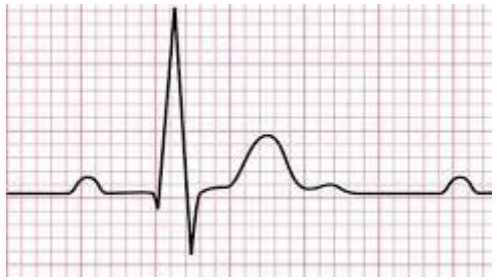
Structured, coded data from linked national healthcare datasets:

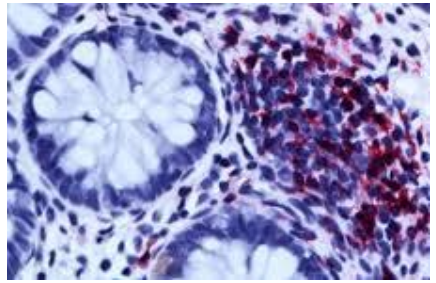
- Can ascertain cases of a wide range of diseases with acceptable accuracy
- Capture only 10-20% of the information from electronic medical records
- Are limited for detailed sub-phenotyping of disease

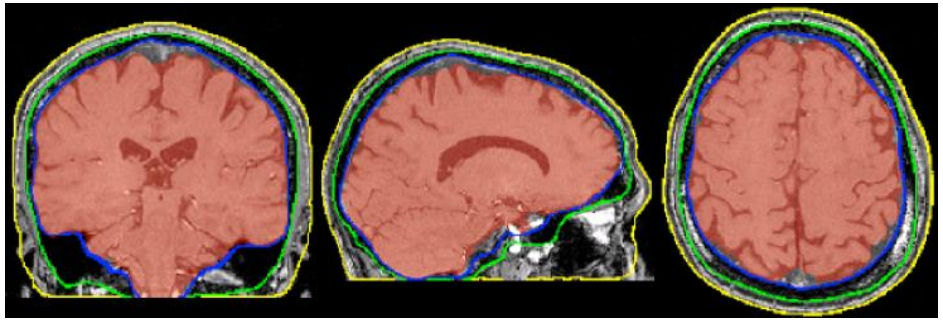
Deeper phenotyping of disease will require multiple unstructured data sources, including:

- Free text of electronic records
- Complex electrical signalling data (ECG's, EEG's etc)
- Histopathology slide sets
- Clinical imaging data


Letter 1
Setting:
Psychiatry letter, April 2015
I was pleased to assess Mr [redacted] during a liaison hospital visit to ward 3, Royal Victoria Hospital on [redacted].4.15. In addition to meeting Mr [redacted], I read his current multidisciplinary team notes, his past psychiatric notes and had the opportunity to discuss nursing management with Charge Nurse, [redacted].
Mr [redacted] has been in contact with my Community Psychiatric Nurse colleagues from my sector Community Mental Health Team for some time and I had received a recent request to offer medical review, although he had subsequently been admitted to the Western General Hospital. I have had discussions with [redacted] (Bridging Team Nurse) and a number of conversations with Mr [redacted]'s wife, who had phoned me directly. Mrs [redacted] had been concerned about his condition







Obtaining these data at national scale is challenging

To extract value from these data on 1000's of outcomes across multiple diseases, we need scalable approaches: crowd sourcing, natural language processing, machine learning, artificial intelligence…

Acknowledgements

MRC Medical Research Council

wellcome

DH Department of Health

British Heart Foundation

CANCER RESEARCH UK

The Scottish Government

Llywodraeth Cynulliad Cymru
Welsh Assembly Government

DIABETES UK
CARE. CONNECT. CAMPAIGN.

biobank uk

Biochemistry analyses in all 500,000 participants

Cardiovascular

- Cholesterol
- Direct LDL-c
- HDL-c
- Triglyceride
- ApoA
- ApoB
- Lp(a)
- CRP

Cancer

- SHBG
- Testosterone
- Oestradiol
- IGF-I

Diabetes

- HbA1c
- Glucose

Renal

- Creatinine
- Cystatin C
- Total protein
- Urea
- Phosphate
- Urate
- Urinary:
  - Creatinine
  - Sodium
  - Potassium
  - Albumin

Bone and joint

- Vitamin D
- Rheumatoid factor
- Alkaline Phosphatase
- Calcium

Liver

- Albumin
- Direct Bilirubin
- Total Bilirubin
- GGT
- ALT
- AST

Note: Haematological assays were conducted during recruitment phase

biobank uk

biobank uk

500,000 participants

22 recruitment centres

89% England

7% Scotland

4% Wales

Industrial scale processes: samples during recruitment

700 participants per day

4,900 sample tubes per day

25,000 aliquots produced per day

15 million 0.85ml aliquots

- Blood
  - whole blood
  - serum
  - plasma
  - red cells
  - buffy coat
- Urine
- Saliva

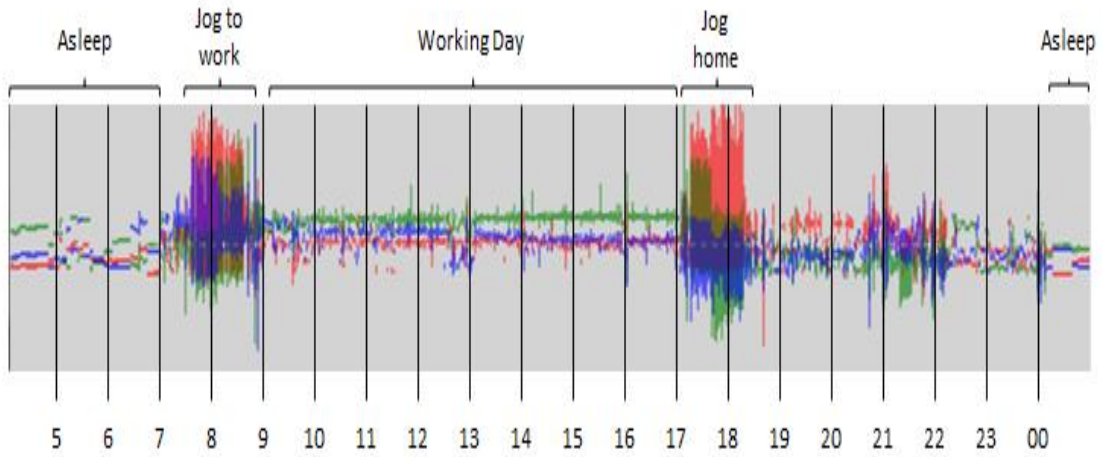
Total > 15 million aliquots

Expected disease cases during follow-up

Condition	2012	2017	2022
Diabetes	10,000	25,000	40,000
Heart attack	7,000	17,000	28,000
Stroke	2,000	5,000	9,000
Chronic obstructive lung disease	3,000	8,000	14,000
Breast cancer	2,500	6,000	10,000
Colorectal cancer	1,500	3,500	7,000
Prostate cancer	1,500	3,500	7,000
Hip fracture	1,000	2,500	6,000
Alzheimer’s	1,000	3,000	9,000

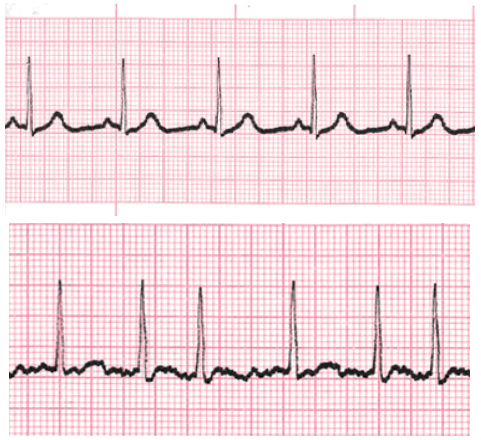
Data from portable wearable devices

Accelerometry data: 100,000 participants

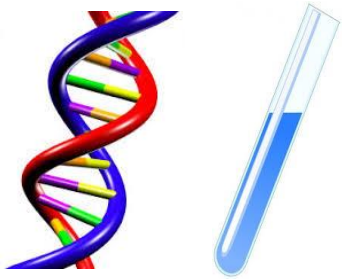


Prospective design and **large size** enable reasonably well-powered studies of (causal) associations between accelerometry and cardiac rhythm measures and later onset disease

Continuous ECG monitoring: 20,000 + participants



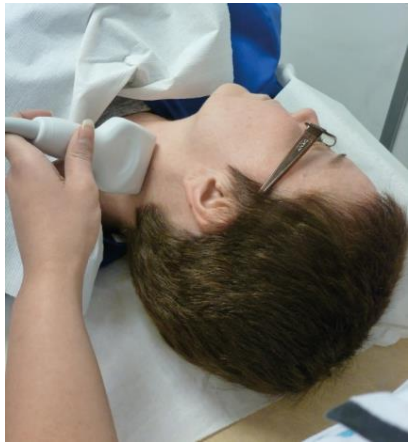
Sample analyses



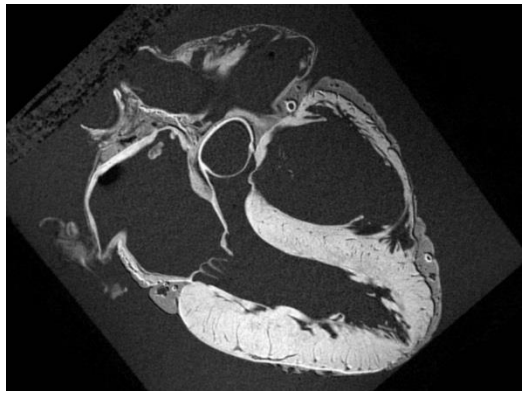
- Genome-wide genotyping of all participants
- Standard panel of assays (e.g. lipids; hormones; metabolic) on samples from all participants
- Exome & whole genome sequencing, proteomics, metabolomics, infectious disease assays, stool microbiome…all underway/planned

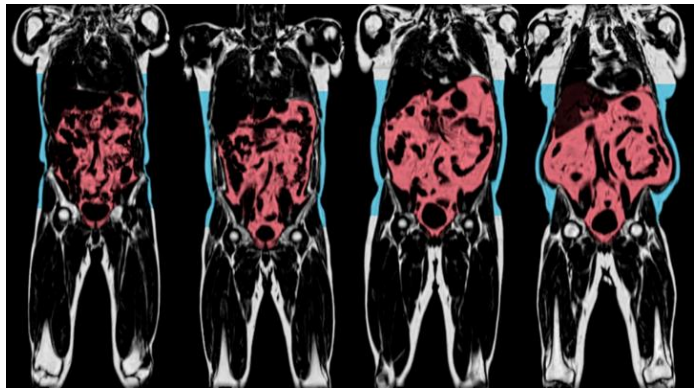
biobank uk

Multimodal imaging of 100,000 participants



>22,000 imaged so far





Prospective design and **large size** enable well-powered studies of (causal) associations between structure and function of organs and later onset disease…but…need **scalable methods** of analysing **complex data** to derive measures for large scale analyses

biobank uk