UK Biobank: research opportunities for the global scientific community

CCACE & Informatics Discussion Day
3 February 2014
Cathie Sudlow
Chief Scientist, UK Biobank
Professor of Neurology and Clinical Epidemiology, University of Edinburgh

What is UK Biobank?

- A **VERY LARGE PROSPECTIVE COHORT STUDY**
- 500,000 UK adults age 40-69 at recruitment
- Baseline data on lifestyle, environment, personal & family medical history, physical measures & biological samples
- Follow-up for disease outcomes over 20+ years
- Establish genetic and environmental determinants of common diseases of middle and old age
- Improve prevention, diagnosis and treatment of cancer, heart disease, stroke, arthritis, dementia….

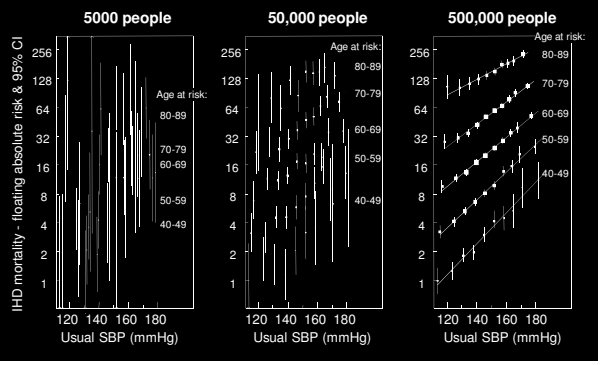
What makes UK Biobank special?

- Very large – will generate sufficient numbers of cases of diseases to allow adequately powered nested case-control and case-cohort studies
- Extensive and detailed exposure measures
- Comprehensive follow-up with careful phenotyping of outcomes
- Open access resource: see www.ukbiobank.ac.uk

Key advantages of prospective studies

- Risk factors can be measured before the disease develops
  (avoid reverse causality & recall bias; improve measurement detail; reduce measurement error)
- Associations can be assessed with a range of diseases
- Appropriate controls can be selected from within the same population as the disease cases

Need for observational studies to be LARGE: IHD versus SBP for 5K *vs* 50K *vs* 500K people in the Prospective Studies Collaboration



Recruitment

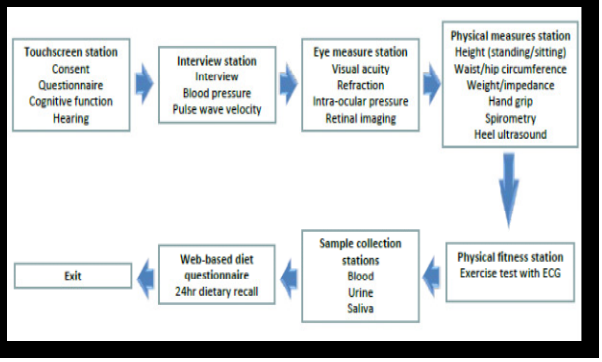
22 recruitment centres in rented office space mainly in large UK cities

503,000 participants recruited 2006-2010

Participant characteristics

- 46% male
- 57% aged 40-59; 43% aged 60-69
- Less socioeconomically deprived than UK average but all strata represented
- 85% urban
- 94.5% white; 5.5% other
- 89% recruited in England; 7% in Scotland; 4% in Wales

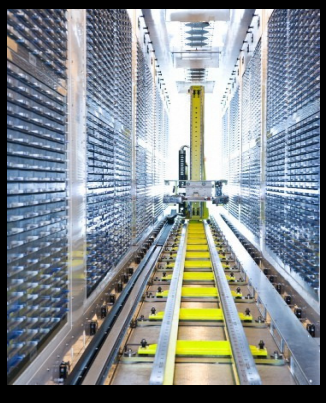
Baseline assessment visit: questions, measures & samples



Automated -80℃ sample archive

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

Total > 15 million aliquots



Prevalent conditions* at recruitment

Condition	Cases (n)
Diabetes	26,000
MI	12,000
COPD	12,000
Stroke	7,000
Breast cancer	11,000
Colorectal cancer	3,000
Prostate cancer	3,000
Rheumatoid arthritis	6,000
Bipolar depression	1400
Schizophrenia	600

** by self report, confirmed by trained interviewer, rounded to nearest 1000*

Incident outcomes* during follow-up

Condition	2012	2017	2022
Diabetes	10,000	25,000	40,000
MI/CHD death	7,000	17,000	28,000
Stroke	2,000	5,000	9,000
COPD	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
Lung cancer	1,000	2,000	4,000
Hip fracture	1,000	2,500	6,000
Alzheimer's	1,000	3,000	9,000

**Estimates based on UK age- and sex-specific rates, adjusted for potential healthy cohort effects and losses to follow-up , rounded to nearest 500*

Neurodegenerative conditions

Condition	Self-report at baseline	Incident outcomes* (n) by the end of:			
	(n)	2012	2017	2022	2027
Stroke	7,000	2,000	5,000	9,000	20,000
Alzheimer's	100	800	3,000	9,000	30,000
Parkinson's	900	1,000	3,000	6,000	14,000

**Estimates based on UK age- and sex-specific rates, adjusted for potential healthy cohort effects and losses to follow-up, rounded to nearest 1000 (or nearest 100 if < 1000)*

Other ongoing enhancements

- Web-based questionnaires for additional detail on exposures & outcomes (cognition, mental health, occupation)
- Repeat of baseline assessment every few years in a subsample of 20-25,000
- Wrist-worn accelerometers being mailed to 100,000 participants to measure physical activity
- Cardiac rhythm monitoring in 100,000
- Standard panel of assays on samples from all participants
- Multimodal imaging in 100,000 - brain, cardiac and abdo MRI, whole body DEXA, & 3D carotid ultrasound
- Genome wide genotyping of all 500,000
- MRC UK Dementia Platform

Long term follow-up in UK Biobank

Value of resource depends on:

- Rich baseline data, samples and further planned enhancements
- Comprehensive and detailed follow-up of health of participants

Key advantages for UK Biobank:

- NHS provides majority of healthcare in the UK
- Cohort-wide linkage to a wide range of routine coded health records possible

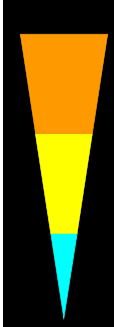
UK-wide linked health care data

Data type	Country	Data provider	Data with UKB	Data in showcase
Certified deaths by cause (ICD codes)	England, Wales	HSCIC	√	√
	Scotland	Central Register & ISD, NSS	√	√
Registered cancers (ICD codes)	England, Wales	HSCIC & NCIN	√	√
	Scotland	Central Register & ISD, NSS	√	√
Hospital episodes (ICD & proccoure codes)	England	HSCIC (HES)	√	√
	Wales	SAIL, Swansea	√	√
	Scotland	ISD (SMR), NSS	√	√
Primary care (Read codes)	England	HSCIC (GPES) or other?	x	x
	Wales	SAIL	√ (50%)	x
	Scotland	HIC, Dundee	√ (50%)	x

Strategy for follow-up

- Key cohort-wide sources for data linkage:
  - Death registrations
  - Cancer registrations
  - Hospital episode data
  - Primary care data
- Questionnaires (web +/- mail)
- Other linkages sought or under discussion:
  - mental health minimum dataset; screening databases; disease registers/audits; imaging; histopathology reports/specimens; NHS dental records
- Future possible linkages:
  - occupational records; DWP ; HMRC; private medical records…
- Close liaison with wide range of data providers
- Working in collaboration with Farr Institute

Staged approaches to outcomes adjudication



APPROACH	CHARACTERISTICS	EXAMPLES
Ascertainment of suspected cases	Cost-effective Feasible Geographically generalisable Scalable	Cause-specific mortality Cancer incidence Hospital discharge records GP records Web self-report q'aire
Confirmation of "caseness"	As above but somewhat higher cost / lower feasibility	Disease registers (eg MINAP) Cross-referencing e-records Targeted blood sampling with cheap assays
Sub-classification of cases	More involved and costly	Targeted blood sampling with costly assays Tumour collection / assays Specialised databases (eg imaging) Review of clinical records

Expert-led adjudication of outcomes

- Outcomes working group
- Developing methods for ascertainment, confirmation, and sub-classification of disease outcomes:

Cancer Diabetes Cardiac outcomes Stroke	Pilots progressing well; preparing for scaled up algorithms and web adjudication
Mental health outcomes Ocular outcomes Neurodegenerative outcomes	Pilots commencing
Respiratory outcomes Musculoskeletal outcomes	Pilots being developed