An Introduction to the UK Biobank

David Brazel

2019-02-18

Overview

The UK Biobank (UKB) is a large prospective cohort of more than 500,000 middle-aged and elderly British men and women. The UKB has quite liberal access policies and is available on application to any "bona fide" researcher for health-related studies. The UKB subjects were deeply phenotyped and genotyped with various follow-up assessments underway, including a death registry and the collection of medical records from the NHS. Some phenotypes were only collected on subsets of the sample (MRI scans or retinal imaging, for example). Most fields are provided without any special justification but restricted fields (e.g., date of birth and precise home coordinates) will be provided "only when absolutely necessary."

Our obligations

- Annual progress report for each project.[1]
- Return of results to the UKB within six months of publication of twelve months of the end of the project.[2] This includes:
  - Derived variables and a description of how they were generated
  - Scripts
  - A copy of the final manuscript
  - A paragraph summarizing the results, to be shared with participants
- If a publication or presentation is "likely to provoke controversy or attract significant public attention," inform the UKB at least two weeks in advance.
- Acknowledge that "this research has been conducted using the UK Biobank Resource" in any presentation or publication and include the project ID number.
- Be sure that you've been added to a particular application before using those data and that your work falls within the scope of the application.
- Remove withdrawn subjects from our files.

[1] `https://bit.ly/2TNcLPb`

[2] `https://bit.ly/2SDD7qu`

Variables

This section will not be an exhaustive survey of the measures available in the UK Biobank. However, I do want to give an overview of

the types of data available and where to find it. Your best bet is the Data Showcase[3], a web app that allows one to browse and search the various data categories and items and provides summary statistics and documentation for each item. When looking at a category or item of interest, be sure to click on the Resources, Notes, and Related Data-Fields tabs. Also, remember that a category can contain both sub-categories and data fields, in which case you'll have to click on the Data-Fields tab to see them.

[3] http://biobank.ctsu.ox.ac.uk/showcase/index.cgi

I will describe the assessment waves in the next section but it is important to know that many fields will have multiple instances. This means that a field was collected at the initial visit but also at subsequent repeat assessment or imaging visits for a subset of participants. This structure can be quite useful for test-retest validity and longitudinal measurement.

Now, let's go through the primary categories.

Population characteristics

This category includes age at recruitment, date of birth, sex, various measures of neighborhood deprivation, and information on subjects lost to follow-up.

UK Biobank Assessment Centre

This category records information gathered from the participants when they visited an assessment centre (this process is explained further in the next section).

The Recruitment subsection has records of the subjects' home location and home area population density, which centre they attended and when, and miscellanous administrative information.

The Touchscreen subsection contains a wide range of data collected through an automated touchscreen questionnaire. Most variables of interest will be somewhere in this section, which I recommend looking through. The routing logic is complex and can be quite abstruse.[4] I recommend making a flowchart to understand the structure of sections of interest.[5]

[4] The questionnaire is comprehensively documented here: https://bit.ly/2GN3qTW

[5] Examples that I made using GraphViz: https://bit.ly/2EfeAiD

The Verbal Interview subsection contains information from an interview conducted by a trained nurse including country of birth, birth weight, employment information, past and current medical conditions, prescription medications taken regularly at the time of the interview, and major surgical operations.

Fields in the Physical Measures subsection include blood pressure and other circulatory measures, a hearing test, eye measures, hand grip strength, body size and composition, bone mineral density, expiratory volume, and electrocardiography.

The Cognitive Function subsection covers various tests of cognitive ability and functioning.

The Imaging category contains both raw data and derived statistics from both whole-body DXA (bone mineral density and body composition) and magnetic resonance imaging of the brain, heart, and abdomen.

The Biological Sampling subsection contains procedural information on the collection of blood, saliva, and urine samples.

The Procedural Metrics section has records of how long each subject took to complete the various stages of the assessment, a timestamp for their completion, and the IDs of the staff associated with each section.

Biological samples

This category contains results from the blood, saliva, and urine assays, including blood count and pathogen antigen titers.

Genomics

These data will be discussed in a later section but I note that this section contains useful metadata which can be viewed in the Data Showcase.

Online follow-up

This category contains results from on-line questionnaires relating to diet, cognitive function, work environment, mental health, and digestive health.

Additional exposures

This category contains information on accelerometer-derived physical activity, residential air and noise pollution, and additional neighborhood information.

Health-related outcomes

This category contains regularly updated information on the participants' health outcomes, largely derived from the NHS. Extensive in-patient hospital records are available but out-patient and primary care data are not. A death register provides date of, age at, and cause of death. A cancer register contains cancer diagnoses, along with the date of the diagnosis and the participant's age.

Recruitment and Assessment Process

Eligible participants were aged 40-69 and lived near one of 22 primary assessment centres[6] in England, Scotland, and Wales.[7] The centre locations were chosen to "provide socioeconomic and ethnic heterogeneity and urban–rural mix" [Sudlow et al., 2015]. Prospective participants were identified from NHS records and other registries. They received an invitation by mail and could obtain more information through a website or by phone. 10% of those contacted enrolled in the study.

The initial visit to an assessment centre occurred between 2006 and 2010. A first visit to an assessment centre consisted of the following stages:[8]

1. Reception and consent
2. Touch screen
   i) Touch screen questionnaire
   ii) Hearing and cognitive tests
3. Interview and blood pressure measurement
   i) Non-medical interview
   ii) First blood pressure measurment
   iii) Medical interview
   iv) Arterial stiffness and second blood pressure measurement
4. Ocular
   i) Visual acuity
   ii) Auto-refraction
   iii) Intra-ocular pressure
   iv) Optic imaging and retinal photography
5. Physical measurements
   i) Hand-grip strength
   ii) Hip and waist size
   iii) Height
   iv) Impedance-based measurement of body composition
   v) Bone density
   vi) Spirometry
6. Electrocardiography
7. Blood, urine, and saliva collection

The procedure varied over the course of sample recruitment, with some measures added or removed. Some measures were applied to only a subset of the sample. For example, accelerometer data were collected from 100,00 participants with 2,500 measured repeatedly. Participants

[6] Edinburgh, Glasgow, Newcastle-upon-Tyne, Middlesbrough, Leeds, Bury, Manchester, Altrincham, Liverpool, Sheffield, Nottingham, Stoke-on-Trent, Birmingham, Oxford, Bristol, Reading, central London, Hounslow, Croydon, Cardiff, Swansea and Wrexham

[7] See `https://www.ukbiobank.ac.uk/all-faqs/` and `https://bit.ly/2Xa9SKy`

[8] Documented here: `https://bit.ly/2S8XhTP`

may, at any time, withdraw from the sample or end further contact.[9] Repeat assessments were conducted for more than 20,000 participants between 2012 and 2013. Imaging visits began in 2014 and are ongoing. At the time of writing, 34,218 participants had been scanned, with a goal of more than 100,000. When participants attended these follow-up visits, they went through the full battery of assessments described above.

[9] 1,299 subjects have been lost to follow-up, see Data-Field 190

Demographics

The UKB is not a representative sample: its participants are older, more female, wealthier, and healthier than the general British population [Fry et al., 2017]. Future versions of this document will expand this section but I will refer you to the cited paper for now.

Genetic Data

Genotyping process

The entire UKB sample was array genotyped but they were not all genotyped on the same array.

Practicalities

Anna Fry, Thomas J Littlejohns, Cathie Sudlow, Nicola Doherty, Ligia Adamska, Tim Sprosen, Rory Collins, and Naomi E Allen. Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants with the General Population. *American Journal of Epidemiology*, 2017. ISSN 1476-6256. DOI: 10.1093/aje/kwx246. URL `http://www.ncbi.nlm.nih.gov/pubmed/28641372`.

Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, Bette Liu, Paul Matthews, Giok Ong, Jill Pell, Alan Silman, Alan Young, Tim Sprosen, Tim Peakman, and Rory Collins. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLOS Medicine*, 12(3):e1001779, March 2015. ISSN 1549-1676. DOI: 10.1371/journal.pmed.1001779. URL `https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1001779`.