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- Remove withdrawn subjects from our files.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

I will expand on this section in a future version but suffice it to say that the UKB is not a representative sample: its participants are older, more female, wealthier, and healthier than Britons at large [Fry et al., 2017].

Genetic Data

Genotyping arrays

The entire UKB sample was array genotyped but not all subjects were genotyped on the same array. Two arrays were used: the UK Biobank Axiom array[10] and the UK BiLEVE array, both of which genotype approximately 800,000 sites. The arrays share more than 95% of their content but, critically, have differential call rates at some sites. The UK BiLEVE array was used for the first 50,000 participants to be genotyped, half of whom were heavy smokers and half of whom were never smokers [Wain et al., 2015]. The other 450,000 participants were genotyped on the Axiom array. The key point is that if you are conducting a genetic study of any trait related to smoking, which includes any trait related to risk tolerance, you need to account for this confounding of phenotype and genotyping methods. One approach, in a GWAS, is to treat the two chips as separate samples.

[10] The design of the array is described here: `https://bit.ly/2BBKfcc` but, broadly, the array is designed to efficiently impute common and low-frequency variants, assess coding variation, and include known or suspected loci.

Data formats

For the full release of genetic data, raw intensities, genotype calls, and imputed genotypes were provided for 488,377 subjects [Bycroft et al., 2017]. The imputed data used a reference panel composed of the HRC and the UK10K, increasing the number of variants from 800,000 to 96 million. For our applications, you are likely to work with the imputed data, which is in BGEN format.[11] Many software packages still have limited or no support for bgen files and it is likely that you will have to convert file formats using the qctool package.

[11] `https://bit.ly/2GtaOzH`

Genotype batch

Genotype calling was performed on groups of samples, called batches. There are a total of 106 batches in the full release. Because assessment centres opened and closed during the recruitment period, the batches are heterogeneous. Statistically significant differences exist on a number of variables, requiring the inclusion of batch as a covariate in genetic association analyses.

Relatedness in the sample

Although the UKB was not designed as a family study, 107,162 pairs are related to the third degree or closer. 179 monozygotic twin pairs, 6,276 parent-offspring pairs, and 22,666 full sibling pairs are present in the dataset. There are also extended family groups, including one remarkable case of eleven individuals with one shared father and eleven different mothers. You will need to either remove related individuals or use methods, such as BOLT-LMM, that accommodate them.

Practicalities

Use the data dictionary

The files for a UKB project will include an HTML data dictionary named ukbNNNN.html where NNNN is the project ID. This file contains an inventory of all data fields included with the project, their IDs (which you need to select variables in your analyses), and hyperlinks to their pages in the Data Showcase. I often reference it while writing UKB analysis scripts.

Treat the data as immutable

As a general best practice, the original data should never be modified. You may, and often should, write scripts that produce quality controlled or otherwise modified versions of the original data but the original files should not be changed.

Genetic data should be kept compressed and indexed

Doing this saves space and processing time. Most software works with compressed files directly, without requiring them to be uncompressed. For tabular genetic data, your best bets are likely to be bgzip for compression and tabix for indexing.

Script everything

Your analyses should never depend on a manual step or on a derived data file that you made while messing around in an R session. In two months, you won't remember what you did or why you did it. Your analyses should be documented and reproducible and that can't happen unless they're scripted.

Put your scripts in version control

Version control software (git is the most popular choice) maintains a record of your changes to your scripts and allows you to go back to an earlier version if necessary. Version control is useful when working alone but essential when collaborating on software with others. Learning to use git takes an hour or two and is very much worth it.[12]

[12] Here's a list of resources: `https://bit.ly/2BF3Qbr`

Use job arrays for easy parallelization

On the cluster, you'll often need to run the same command, say a QC step, on many files, perhaps one VCF file per chromosome. Slurm job arrays make this dead simple, as explained here.

If possible, use notebooks

Ideally, our exploratory work is also documented. The easiest way to do this is with a "notebook," a tool that allows you to weave together prose, code, equations, and graphics into a single interactive document. R, Python (Jupyter Notebook), and Mathematica are examples of languages that support notebooks.

References

Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Olivier Delaneau, Jared O Connell, Adrian Cortes, and Samantha Welsh. Genome-wide genetic data on ~500,000 UK Biobank participants. *bioRxiv*, 2017. DOI: http://dx.doi.org/10.1101/166298.

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`.

Louise V Wain, Nick Shrine, Suzanne Miller, Victoria E Jackson, Ioanna Ntalla, María Soler Artigas, Charlotte K Billington, Abdul Kader Kheirallah, Richard Allen, James P Cook, Kelly Probert, Ma'en Obeidat, Yohan Bossé, Ke Hao, Dirkje S Postma, Peter D Paré, Adaikalavan Ramasamy, Reedik Mägi, Evelin Mihailov, Eva Reinmaa, Erik Melén, Jared O'Connell, Eleni Frangou, Olivier Delaneau, Colin Freeman, Desislava Petkova, Mark McCarthy, Ian Sayers, Panos Deloukas, Richard Hubbard, Ian Pavord, Anna L Hansell, Neil C Thomson, Eleftheria Zeggini, Andrew P Morris, Jonathan Marchini, David P Strachan, Martin D Tobin, and Ian P Hall. Novel insights into the genetics of smoking behaviour, lung function, and chronic obstructive pulmonary disease (UK BiLEVE): a genetic association study in UK Biobank. *The Lancet Respiratory Medicine*, 3(10):769–781, October 2015. ISSN 22132600. DOI: 10.1016/S2213-2600(15)00283-0. URL `http://linkinghub.elsevier.com/retrieve/pii/S2213260015002830`.