UK Biobank

A Data Management Plan created using DMPonline

Creators: Sander W. van der Laan, Vinicius Tragante, Kristel Kool, PLEASE UPDATE YOUR DETAILS, Charlotte Onland, Jessica van Setten, Vinicius Tragante

Affiliation: Other

Funder: European Commission

Template: UMC Utrecht DMP

ORCID iD: 0000-0002-8223-8957

ORCID iD: 0000-0002-4934-7510

ORCID iD: 0000-0002-2360-913X

ORCID iD: 0000-0002-1692-8669

ORCID iD: 0000-0001-6888-1404

ID: 80411

Start date: 01-01-2021

End date: 01-01-3000

Last modified: 06-01-2022

UK Biobank

1. General features

1.1. Please fill in the table below. When not applicable (yet), please fill in N/A.

DMP template version	29 (don't change)
ABR number *(only for human-related research)*	n/a
METC number *(only for human-related research)*	C-01.18
DEC number *(only for animal-related research)*	n/a
Acronym/short study title	UK Biobank
Name Research Folder	
Name Division	Laboratories, Pharmacy, and Biomedical genetics
Name Department	Central Diagnostic Laboratory
Partner Organization	
Start date study	2021-01-01
Planned end date study	3000-01-01
Name of datamanager consulted*	Saskia Haitjema
Check date by datamanager	January 6th 2022

1.2 Select the specifics that are applicable for your research.

- Retrospective study
- Observational study
- Use of Questionnaires
- WMO
- Fundamental / translational study

We will use data from the UK Biobank (UKB, https://www.ukbiobank.ac.uk) as reference in many studies or as part of a course curriculum in practicals to learn genetic analyses methods, as well as in study projects of researchers in different UMC groups.

UK Biobank is a large-scale biomedical database and research resource, containing in-depth genetic and health information from half a million UK participants. The database is regularly augmented with additional data and is globally accessible to approved researchers undertaking vital research into the most common and life-threatening diseases. It is a major contributor to the advancement of modern medicine and treatment and has enabled several scientific discoveries that improve human health.

2. Data Collection

2.1 Give a short description of the research data.

Subjects	Volume	Data Source	Data Capture Tool	File Type	Format	Storage space
Human	1	Genotype data	R, SNPTEST, GCTA, etc	PLINK-format, Oxford-format	.vcf, .bed/.bim/.fam, .gen/.sample	±15Tb
Human	1	Phenotype data	R, python, etc	flat text file	.tab	±50gb

2.2 Do you reuse existing data?

- Yes, please specify

Existing data from the UKB:
- Genotype data
- 'Clinical' data, i.e. extensive phenotype information, see below for more information

Discover UK Biobank
UK Biobank is a world-leading biomedical database to enable scientific discoveries that improve human health. Our goal is to inspire the imaginations of health researchers around the world to meet the challenge of greater understanding, prevention, and treatment of a range of serious illnesses. UK Biobank has Research Tissue Bank Approval till 2026 and through access to an unmatched amount of biological and medical data on half a million people living in the UK, we can enable your vision of improving human health.

1. UK Biobank is a longitudinal study; it follows the health of 500,000 volunteer participants.
2. Participants were aged between 40-69 years when they joined UK Biobank between 2006-2010.
3. Each participant attended a baseline assessment at a centre in England (89%), Scotland (7%) and Wales (4%).
4. Participants provided their consent for long-term follow-up.
5. Participants answered lots of questions about their health & lifestyle.
6. Participants donated samples of blood, urine and saliva for long-term storage and analysis.
7. Physical measurements were also taken (e.g. height, weight, spirometry, blood pressure, heel bone density).
8. Many participants have undertaken MR brain & heart imaging, activity monitoring and online follow-up questionnaires.
9. We have genetic data on all 500,000 participants.
10. UK Biobank is not representative of the general population with evidence of a ‘healthy volunteer’ selection bias, details of which are published online.

2.3 Describe who will have access to which data during your study.

Please note, that the data has been de-identified for the purpose of public sharing.

Type of data	Who has access
Pseudonymized data	Research team, Datamanager

2.4 Describe how you will take care of good data quality.

#	Question	Yes	No	N/A
1.	Do you use a certified Data Capture Tool or Electronic Lab Notebook?			x
2.	Have you built in skips and validation checks?			x
3.	Do you perform repeated measurements?			x
4.	Are your devices calibrated?			x
5.	Are your data (partially) checked by others (4 eyes principle)?			x
6.	Are your data fully up to date?	x		
7.	Do you lock your raw data (frozen dataset)	x		
8.	Do you keep a logging (audit trail) of all changes?	x		
9.	Do you have a policy for handling missing data?			x
10.	Do you have a policy for handling outliers?	x		

2.5 Specify data management costs and how you plan to cover these costs.

#	Type of costs	Division ("overhead")	Funder	Other (specify)
1.	Archiving	x		
2.	Storage	x		
3.	Maintenance Dataset		x	
4.	Datamanager	x		
5.	Data analysis tool	x		

2.6 State how ownership of the data and intellectual property rights (IPR) to the data will be managed, and which agreements will be or are made.

UK Biobank is a large-scale biomedical database and research resource that is enabling new scientific discoveries to be made that improve public health. The resource provides accredited researchers access to medical and genetic data from half a million volunteer participants to improve our understanding of the prevention, diagnosis and treatment of a wide range of serious and life-threatening illnesses.

Through the long-term commitment of 500,000 participants, together with the support of our funders, we are enabling new scientific discoveries that benefit people's health.

Development of UK Biobank involves engagement with our funders and extensive consultation with the public and scientific community. We seek to implement scientifically rigorous processes on a very large-scale, that are ethically robust and ensure we achieve our aim to enable novel scientific discoveries. UK Biobank receives direction from the UK Biobank Board and support from a range of committees and expert advisory groups.

Thus the UK Biobank is the main manager of the data and the study collection.

3. Personal data (Data Protection Impact Assessment (DPIA) light)

Will you be using personal data (direct or indirect identifying) from the Electronic Patient Dossier (EPD), DNA, body material, images or any other form of personal data?

- No, go to 4.1

4. Data Storage and Backup

4.1 Describe where you will store your data and documentation during the research.

The digital files will be stored in a secured Research Folder Structure of the UMC Utrecht. We will need 15+ Tb storage space, so the capacity of the network drive will be sufficient.

For purposes of analyses digital files are partly and temporarily stored on the high-performance computer cluster (HPC) facilitated by the institute or a UMC Utrecht owned and managed device.

Data storage is only accessible to authorized personnel.

Phenotype data is accessible based on an application number as given out by the UK Biobank as part of an approved application. The phenotype data should be stored per-group and only made accessible to named in the respective application. These are not stored in the RFS associated with this DMP.

4.2 Describe your backup strategy or the automated backup strategy of your storage locations.

All (research) data is stored on UMC Utrecht networked drives from which backups are made automatically twice a day by the division IT (dIT).

We will have multiple copies 1) at the HPC, and 2) at the UMC internal network.

5. Metadata and Documentation

5.1 Describe the metadata that you will collect and which standards you use.

We do not collect anything else, but the data we can obtain through a download.

5.2 Describe your version control and file naming standards.

We will use GitHub as version control with a specific GitHub repository for the each individual project.
We will use the release-system native of GitHub and where possible link it to Zenodo (code only!).

6. Data Analysis

6 Describe how you will make the data analysis procedure insightful for peers.

We will write an analysis plan in which we state why we will use which data and which statistical analysis we plan to do in which software. The analysis plan will be stored at GitHub or potentially through a pre-registration server, e.g. OSF. This way this will be findable for our peers.

7. Data Preservation and Archiving

7.1 Describe which data and documents are needed to reproduce your findings.

The data package will contain: the raw data, the study protocol describing the methods and materials, the script to process the data, the scripts leading to tables and figures in the publication, a codebook with explanations on the variable names, and a 'read_me.txt' file with an overview of files included and their content and use.
Where it is relevant this is amended by an Electronic Lab Notebook (ELN) and handwritten (legacy) lab journals.
After finishing the project, documentation for the ELN will be stored at the UMC Utrecht [GIVE FULL PATH] and is under the responsibility of the Principal Investigator of the research group.
** I will update 'XXX' in this answer when available.*

7.2 Describe for how long the data and documents needed for reproducibility will be available.

Data and documentation needed to reproduce findings from this WMO study will be stored for at least 10 years.

7.3 Describe which archive or repository (include the link!) you will use for long-term archiving of your data and whether the repository is certified.

We do not 'own' the data, it is controlled/managed by the UK Biobank. We will only keep copies for local use, and potentially archive projects through Archivemetica and share codes used

publications etc through DataverseNL according to the principles of FAIR. At the same time a copy will remain at the department server in the existing Research Folder Structure and is under the responsibility of the Principal Investigator of the research group.

7.4 Give the Persistent Identifier (PID) that you will use as a permanent link to your published dataset.

When we get DOI-codes we will update this plan to included these.

8. Data Sharing Statement

8.1 Describe what reuse of your research data you intend or foresee, and what audience will be interested in your data.

Specifically the methods and codes developed for the use of this data will be of interest to our peers. Since the data is managed by the UK Biobank we refrain from stating anything regarding data re-use, other than that in general these data make for an excellent population reference for multiple purposes.

8.2 Are there any reasons to make part of the data NOT publicly available or to restrict access to the data once made publicly available?

- Yes (please specify)

As the data is privacy-sensitive, and managed by theUK Biobank we will refrain from sharing these data publicly; this should go through UK Biobank.

8.3 Describe which metadata will be available with the data and what methods or software tools are needed to reuse the data.

Publications will be open access. The study protocol and this Data Management Plan will also be available.
Along with the publication, the codebook of the data and scripts of analyses will be available through GitHub.
Data (raw or processed) will be accessible under conditions set forward by theUK Biobank.

8.4 Describe when and for how long the (meta)data will be available for reuse

- Other (please specify)

Meta data will be accessible under conditions set forward by theUK Biobank.

8.5 Describe where you will make your data findable and available to others.

We will publish and archive publication, codes, etc as described above through Archivemetica (local archiving) and DataverseNL (public) with a note that the data will be accessible under conditions set forward by the UK Biobank.