UK Biobank / Data Management & Sharing Plan

Managing data access

1. What types of data will UK Biobank hold? What are the principal data standards that are being used for the data that UK Biobank collects and manages?

   A. UK Biobank is designed to hold data of any sort on participants, ranging from simple demographic information to detailed clinical data such as ECG results and eye scans. Wherever possible these data are and will be stored in widely used open formats (for instance SOC2003[1] coding for occupations, DICOM[2] for MRI) however no single standard covers the diverse range of information held. All encodings used are fully documented and available to researchers. Each coding is shown in the Showcase database alongside the field it relates to. Further, any researcher receiving data also receives a list of the relevant codings for the data they are sent.

   What restrictions are there in the MTA on researchers disclosing individual's genotypes when publishing results?

   A. A researcher is not entitled to publish any material which could lead to the identification (inadvertent or otherwise) of an individual. This protocol is set out in more detail in our data de-identification protocol.

   The UK Biobank Board data access committee is a sub-committee of their management board, and therefore not unequivocally impartial if there is a dispute over an application?

   A. As the Board of UK Biobank is the body ultimately responsible for access - the Access Sub-Committee (ASC) is a sub-committee of the Board - it is appropriate that the Board should retain control over the adjudication of any disputes. That said, there is nothing to prevent independent experts being consulted by the ASC (or the Board).

   Is it clear that UK Biobank personnel wanting to undertake own-account research on the data will undergo the same access procedure?

   A. UK Biobank itself won't be conducting research. Members of the UK Biobank staff – including members of the ASC, the CEO/PI or the Chief Scientist – and the institutions to which they are associated might well seek access. This would be adjudicated in the normal way: in other words they would have to go through the Access Procedures like any other applicant and their applications would be reviewed independently by the ASC (any ASC member whose institution is involved in an application is not entitled to participate in the decision as to whether or not to grant access).

2. What types of data will UK Biobank make available to researchers? And what types of data will not be made available?

---

1 Data standard for "occupations".

2 The Digital Imaging and Communication Standards In Medicine.

A. *There is no general restriction on the types of data (which UK Biobank holds) that it will make available to researchers, with the exception of items (such as NHS number, name and postal address) that could be used to identify individual participants. That said:*

UK Biobank may restrict the amount of data which it makes available to researchers, either on the basis that the request is excessive[3] *or that in light of the size of the data sets (eye image data for example) it may be appropriate to make this available to researchers in other forms / formats*[4]*;*

UK Biobank might decide to restrict access to certain types of data if it had concerns about its accuracy or its provenance. In general, however, UK Biobank will not impose its own quality criteria on the data, but rather will describe the origin of the data and the methods of data collection so that the data quality can be judged (by the researcher) on the basis of the particular research question being addressed; and

It may be the case – although UK Biobank is not currently aware of any specific instances - that certain of the linkage data which UK Biobank incorporates within its database may only be provided with certain restrictions attached (which the relevant third party provider requires[5]*).*

3. What type of information about the data (i.e. metadata) will be made available to those wanting to access the resource? How will researchers be able to select the data that they require? What data standards are being used for the metadata?

A. *UK Biobank's Data Showcase will display on a public website all the data types which are available within the Resource in a grouped format (i.e. a univariate distribution of the variable in appropriate categories, not at the individual participant level). Further, there will be certain limitations on the amount of publicly available information on certain categories of sensitive data entries (for example the category "number of sexual partners*[6]*").*

For clarification, we would add that there is a clear distinction between (1) 'metadata' (or summary data) on participant characteristics, exposures and outcomes, which are displayed in the Data Showcase to anyone who visits the website and (2) data on individual, anonymised participants which will be provided only to approved bona fide researchers who have demonstrated that the particular research use the data will be put to is in the public good (and that they fulfil all the other requirements described in the Access Procedures).

---

3 "Excessive" will be assessed on a case-by-case basis as it depends on the research question in issue. Whether a request is excessive is a matter of context. UK Biobank is not trying to prevent researchers accessing large datasets – genetic sequence data for example – rather trying to par down researchers from asking for data they don't need.

4 For example, image data is of such a size that downloading will not be practical and instead the data will accessed on UKB's servers.

5 For example, requirements which may be imposed on UK Biobank from data linkage sources, such as GPES.

6 These data will be available to bona fide researchers, but univariate distributions of sensitive data will not be shown in Showcase.

Detailed information about each data field will be made available in the Data Showcase[7] and in explanatory documents, which provide the necessary background as to how the measures were taken. Researchers will be able to select the type and parameters for the data they require through UK Biobank's Application Management System ("UK Biobank's AMS").

The metadata which describe the contents of the data repository are held in a standard relational database, which is accessible to both researchers and the general public through categorised and searchable web interfaces. As above, no one standard[8] for metadata has been used as the data cover a number of different categories. UK Biobank would be prepared to work with researchers to develop and produce new encodings for data, which will be treated as a data assay (see section below).

What is the meaning of "reverse-anonymised form encoded using an irreversible algorithm."

A. The anonymisation of the data is irreversible in the hands of the researchers by way of the irreversible algorithm. UK Biobank retains an internal database which enables UK Biobank (and only UK Biobank) to reverse anonymise the data.

4. By what methods will researchers be able to access the data that they require?

A. By using UK Biobank's AMS which requires the researcher to meet the requirements of UK Biobank's Access Procedures and enter into a (standardised) material transfer agreement ("UK Biobank's MTA") with UK Biobank.

Researchers whose applications are approved will be able to download the encrypted datasets that they have requested at Application (or in the case of very large datasets, access them on UK Biobank's servers); these data sets are individually password-protected, with passwords being sent separately. UK Biobank may undertake sample analysis and will then provide the data and the assay results to researchers. UK Biobank will not generally undertake data analysis on behalf of researchers, but it may undertake certain data integration tasks on behalf of researchers – most likely in relation to their use of the results of other researchers analyses of UK Biobank data– and these will be treated in the same way as sample assays.

5. How will the research community be made aware of the availability of datasets from UK Biobank?

A. Principally through UK Biobank's website and periodic updates, as well as through the usual public channels such as published articles (e.g. a Lancet article was timed to co-incide with the initial launch of the Resource).

Specifically in relation to the UK Biobank website, the website will highlight changes and developments, such as the release of new data types, the establishment of data linkage and the status of enhancement proposals. It is also planned to update researchers by email with news about the development of items of interest to them.

---

7 These data fields are distinct to UK Biobank and thus the data fields are bespoke.

8 For example DDI-3 is more of social-science standard.

Further, UK Biobank will advertise the existence and content of returned datasets via the Data Showcase website and in due course using email alerts.

Data generated by UK Biobank and researchers

6. What categories of data will be generated during a research project?

A. In the course of a research project the following broad categories of data may be generated (either by UK Biobank or by the researcher) from the UK Biobank samples or the UK Biobank data:

- *Assay Data: these are data generated from analyses of the samples*[9]*. These assays may be conducted by or on behalf of UK Biobank or by the researcher themselves. Assay Data includes by way of example biochemical analyses (lipids such as HDL or LDL cholesterol) and genetic sequence data;*

- *Analysis Data: these are data generated from analyses of other UK Biobank materials or UK Biobank data. As above, these assays may be conducted by or on behalf of UK Biobank or by the researchers themselves. Analysis Data include by way of example, data generated from the analysis of eye (and other) images, spirometry, EGCs, activity monitoring, diet questionnaires and similar;*

- *Derived Data Variables: these are data generated from analyses or functional combinations of the UK Biobank data or Assay data. An example of a Derived Data Variable includes Body Mass Index (BMI), being a function of height and weight;*

- *Researcher Analyses: these are methodologies, applications and mathematical devices used by researchers, such as power calculations, algorithms and statistical correlations, used by the researcher;*

- *Researcher Results (which includes Research Results Data): these are respectively the qualitative conclusions of the Researcher and the quantitative analyses which underpin those conclusions. These data will go back into the resource and will be carefully evaluated (for example ECG data) and its provenance highlighted.*

7. What's the position regarding ownership[10] of these categories of data?

A. *In terms of "ownership" UK Biobank is the "owner*[11]*" of:*

- *the UK Biobank samples and the UK Biobank data;*

---

9 Assays commonly qualitatively assess or quantitatively measure the presence, amount and/or functional activity of a particular analyte.

10 Legally, we do own the samples and we do own the data collection. It is considered preferable to refer to the actual situation and then try to soften it – rather than vice versa.

11 "Guardian" or "Custodian" is probably a more appropriate and less legalistic term than "owner".

- *any Assay Data;*

- *any Analysis Data; and*

- *any Derived Data Variables.*

The researcher is entitled to use all these data for the duration and purposes of the research project and to the extent necessary to support any filings for IPRs.

Other than as part of its Researcher Results, the researcher is not entitled to republish or otherwise make available any UK Biobank Data, any Assay Data, any Analysis Data or any Derived Data Variables at the individual participant level.

To the extent that this has not already been done, at the conclusion of the research project, the researcher must provide all Assay Data, Analysis Data and Derived Data Variables to UK Biobank.

The researcher owns the Researcher Analyses and Researcher Results, subject to providing such information back to UK Biobank which UK Biobank may in turn make available to other (approved) researchers.

Data returned to Biobank

8. When will researchers be required to return research results and supporting raw data to the Resource?

A. Please refer to Annexe II of the Access Procedures and to clause 6.3 of UK Biobank's MTA. In summary all researchers are required to return their results and supporting data - irrespective of whether such results and data do or do not support the researcher's hypothesis - to UK Biobank. What this means in practice will vary from one project to another but in summary it will likely include the data set used, the methodology, any assay data[12] and the derived data. The objective is that UK Biobank can be in a position to make the research results and supporting data available to other researchers in such manner that the analyses can be peer reviewed through a process of replication.

What are the provisions (if any) for when researchers return results and supporting data that are augmented with results and data from additional non-UK Biobank individuals?

A. Researchers are not obliged to return results for non UK Biobank individuals. Should they do so, the data will be kept as part of the discrete researcher results but they will not be incorporated into the UK Biobank resource.

9. What information will you be supplying to research users of Biobank to ensure that their data returns are in a suitable format to be compatible with Biobank's systems?

12 Including data from sample and data assays and any sequencing.

A. *The format for data return will be developed over the coming months, however it is expected to be a tabular and/or xml specification[13] referencing new data items (either derived data fields created from analysis of existing data or new data fields related to assay results) against the anonymised IDs supplied to the researcher. Details will vary according to the type of data being returned, which will be diverse – for instance simple numbers or three-dimensional retinal mappings.*

10. How will you control the quality of data returned to UK Biobank's database?

A. *Researchers will return their results data to UK Biobank. These files of research results (which will be clearly marked for provenance) will in turn be available to other researchers who apply to use the UK Biobank resource. UK Biobank does not propose to audit these data for quality control: instead the fact that they are available for review by other researchers should result in an effective form of peer review. UK Biobank will, however, check the results and data to ensure that what has been provided is effectively a "complete set".*

In the event that data are included within UK Biobank's own database (such as assay data or sequence data) the origin of such data and the means by which they were derived will be clearly flagged. As noted above, although UK Biobank will require a reasonable degree of confidence in data to be incorporated within its database, it does not propose to specifically audit such data for quality control.

It should also be noted that there is a specific obligation to return all results (including negative results) to UK Biobank. Finally, there is a requirement to credit UK Biobank in the following form:

- *"This research has been conducted using the UK Biobank Resource."; or*
- *where the researcher has used the results and data generated by other researchers as distinct from data within UK Biobank "This research has been conducted using results and data generated by previous researchers who have used the UK Biobank Resource."; or*
- *where there is combination "This research has been conducted using both the UK Biobank Resource and also using results and data generated by previous researchers who have used the UK Biobank Resource."*

11. What data formats and quality standards will be applied to enable the returned data to be shared effectively?

A. *Data incorporated directly in the central repository will be clearly marked (for origin/provenance) and held in the same relational format as the UKB-produced information, and documented with meta-data in the same fashion. Researchers will be required to provide a clear description of the methodology through which their data were derived.*

13 More specifically, UK Biobank will require that derived data be returned in a tabular format, stored as ascii text using either character separated fields or XML. Complex data (e.g. images) will be returned as a collection of individual files with referenced by a tabular index containing descriptions and appropriate checksums.

The checks/procedures required to verify data have not yet been formalised and may vary on a case-by-case basis according to the trustworthiness/reputation of the data supplier.

In terms of sanction, there are a number of legal remedies available to UK Biobank to "ensure" compliance but the most effective will probably be imposing restrictions (or a complete ban) on using the resource again by the researcher and, potentially their Institution, and reporting non-compliance to sponsoring or funding bodies.

12. How will researchers who are not direct users of the cohort be able to access the data that are returned to UK Biobank?

A. Any researcher may apply to access the results of other researchers by going through UK Biobank's Access Procedures and entering into UK Biobank's MTA.

Data preservation

13. What requirements or information on data preservation, for example whether researchers should keep their records and for how long, will be given to researchers?

A. As UK Biobank will retain a copy of the results data, it will not (and does not need to) proscribe a time limit to the researchers for the retention of data[14]*.*

Data security

14. What data security standards will be applied to UK Biobank data?

A. Only a very limited number of UK Biobank staff – and only those who need to use such technical facility as part of their operational responsibilities – have access to the systems which could be used to identify any participant. UK Biobank is working in accordance with ISO 27000 and its systems are subject to regular internal and external audit.

All researchers will be obliged to complete certain fields in the UK Biobank Access Procedures setting out their data security systems and protocols. All data supplied by UK Biobank to researchers will be in reverse-anonymised form, encoded using an irreversible algorithm. All researchers will be obliged to complete certain fields in the UK Biobank Access Procedures setting out their data security systems and protocols. Further, UK Biobank has the right (under the MTA) to audit the security systems of any researcher. That said, once the data is released to a researcher – who has signed an MTA – there is a limit as to how far (practically speaking) UK Biobank can oversee its further dissemination

15. Will the same standards be required of those using UK Biobank data?

A. Researchers will be obliged to maintain the appropriate standards (set out in clause 4 of the MTA) regarding their storage and use of UK Biobank data.

[14] As it very difficult to "destroy" data – it gets shifted from one part of computer's memory to another – UK Biobank will require that they render it inaccessible for further use.

Other relevant repositories

16. How will UK Biobank link its data with other datasets to enhance its value?

A. UK Biobank will link to the health-related records of the participants (through cancer and death registries and through hospital episode and primary care data) and in time to a range of other health-related records (further details are available on UK Biobank's website).

17. Have you considered whether researchers or UK Biobank are responsible for sharing the data with other recognised, relevant repositories?

A. The researcher is responsible for returning the results data to UK Biobank and in turn UK Biobank will make these data available to other researchers (who apply to use the UK Biobank Resource through UK Biobank's Access Procedures).

The possibility of lodging some subsets of the data (e.g. genomes) with other data repositories such as EGA needs to be addressed.

A. As a general rule, researchers are not entitled to "share" data with other researchers without the other researchers registering – which is not a particularly onerous task – with UK Biobank as part of the access procedures.

We do recognise that there are issues with genetic datasets and we are considering these matters in detail and will hope to have a protocol sometime in 2013 (although it will in large part depend on what genotyping activity is undertaken or funded).

Resources

18. What staff and resources will you allocate to implement this plan?

A. UK Biobank has installed a dedicated access team within its Cheadle Co-ordinating Centre to handle the administrative side of all access applications. The supervisory part of the process will be handled by UK Biobank's Access Sub-Committee and other relevant parties, such as Ethox (whose roles and remits are set out in the Access Procedures). The Clinical Trial Service Unit, University of Oxford (under contract with UK Biobank) provides dedicated statistical/epidemiological expertise along with IT development/support and hosting/management infrastructure.

Further, UK Biobank will charge researchers on a cost recovery basis. These costs will cover administrative and other staffing costs, assay costs, retrieval / delivery costs: indeed any costs that UK Biobank incurs in the application process. These costs will be charged at UK Biobank's own internal cost or at the third party cost to UK Biobank.

The two slides below show the summary components and the review process.

System components

Process overview