UK Biobank Showcase User Guide: Getting Started

1 Introduction

UK Biobank holds an unprecedented amount of data on half a million participants aged 40-69 years recruited between 2006 and 2010 throughout the UK. Showcase (available through http://www.ukbiobank.ac.uk) aims to present the data available for health-related research in a comprehensive and concise way, and to provide technical information for researchers considering applying to use the resource.

This user guide is designed to give you an overview of the data and provides some instructions on how to navigate your way through the system.

Suggestions and information for new users:

- **Have a printout of this user guide handy when you first use Showcase**
- **Read the background information about UK Biobank and details on access procedures at http://www.ukbioank.ac.uk**
- **Take time to familiarize yourself with the Showcase structure, the accompanying documentation and the descriptions provided for each data-field before completing a preliminary application to use the resource.**
- **Note that Showcase is still under development. We expect to improve and refine it over the months ahead, and to add further data as they become available (see www.ukbiobank.ac.uk/timetable/).**

If you encounter problems or faults, please email showcase@ukbiobank.ac.uk

Improving the health of future generations

2 Data included in UK Biobank

2.1 Data collected at the Assessment Centre

All participants in UK Biobank were recruited through assessment centres, designed specifically for this purpose. Data collected at the assessment visit included information on a participant's health and lifestyle, hearing and cognitive function, collected through a touchscreen questionnaire and brief verbal interview. A range of physical measurements were also performed, and which included: blood pressure; arterial stiffness; eye measures (visual acuity, refractometry, intraocular pressure, optical coherence tomography); body composition measures (including impedance); hand-grip strength; ultrasound bone densitometry; spirometry; and an exercise/fitness test with ECG. Samples of blood, urine and saliva were also collected.

During 2006, over 3,000 participants were included in the pilot phase of recruitment. Where possible, data collected from the pilot and the main recruitment phases have been combined (note: at the time of release, this work is still ongoing). Where modifications to the protocol were made after the pilot study, the data-fields from the pilot and main recruitment phase are listed separately (e.g., touchscreen questions on medications and family history). In addition, cognitive function tests that were felt to be too time-consuming and/or relatively uninformative were omitted from the main phase of recruitment (i.e. the light memory test' on the touchscreen questionnaire and the 'word test' that was performed during the verbal interview stage).

2.2 Other data

These include (now or in the future - see http://www.ukbiobank.ac.uk/timetable/):

- Data on the biological samples held in UK Biobank's laboratory store;
- Additional exposure data not collected at the assessment visit (e.g., data from web-based dietary questionnaires);
- Data on health-related outcomes via linkage to a range of health-related records.

Some data categories currently have no data fields, but are included in the Showcase to indicate that data are forthcoming.

3 Finding data in Showcase

You can find data through two main routes:

BROWSE: Use this to navigate your way through hierarchical categories and subcategories of interest to data-fields (i.e. variables) of interest. **This will be the most appropriate tool for most researchers wishing to find and select data for their application to use the Resource.**

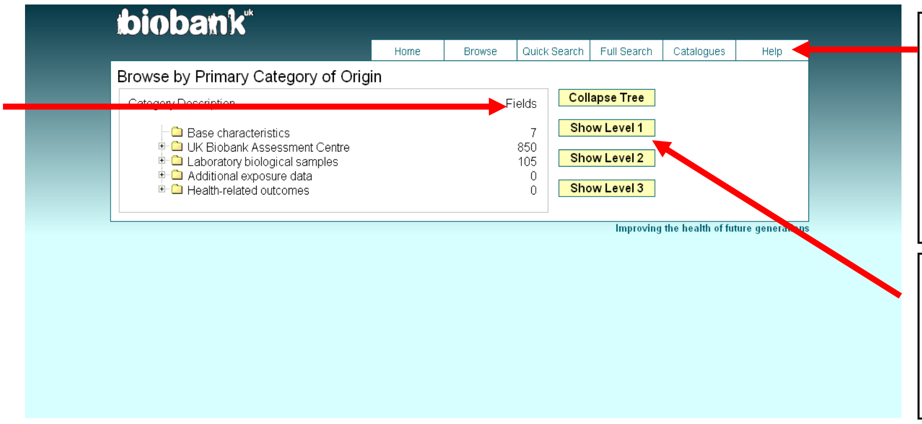
SEARCH: This is based on a text search of the data-field name and its notes, and uses the Boolean operators ‘**&**’ and ‘**|**’ to denote the ‘AND’ and ‘OR’ functions. By default, only whole word matches are returned, although you can use the asterisk ‘*****’ as a wild-card character at the beginning or end of text to search for words containing that text. Please see the **HELP** page on ‘Searching text’ for more details. The **Full Search** facility allows you to conduct a search using specific criteria based on the type of data-field (see Section 5 for more details).

A full list of data-fields, categories and documents can be found in **CATALOGUES.**

4 Data categories and sub-categories

Data are organized in a tree structure, accessible via **BROWSE**, with the main categories based on the origin of data collection (Figure 1). These include ‘Base characteristics’ (some general characteristics of participants known before arrival), ‘UK Biobank Assessment Centre’ (data obtained at the Assessment Centre), ‘Laboratory biological samples’ (data on biological samples), ‘Additional exposure data’ (data collected outside the Assessment Centre), and ‘Health-related outcomes’ (data from linkage of participants to health-related records). Please see the **HELP** page on ‘Browse’ for more details.

The **Fields** column lists the number of data-fields in each category (and its sub-categories)



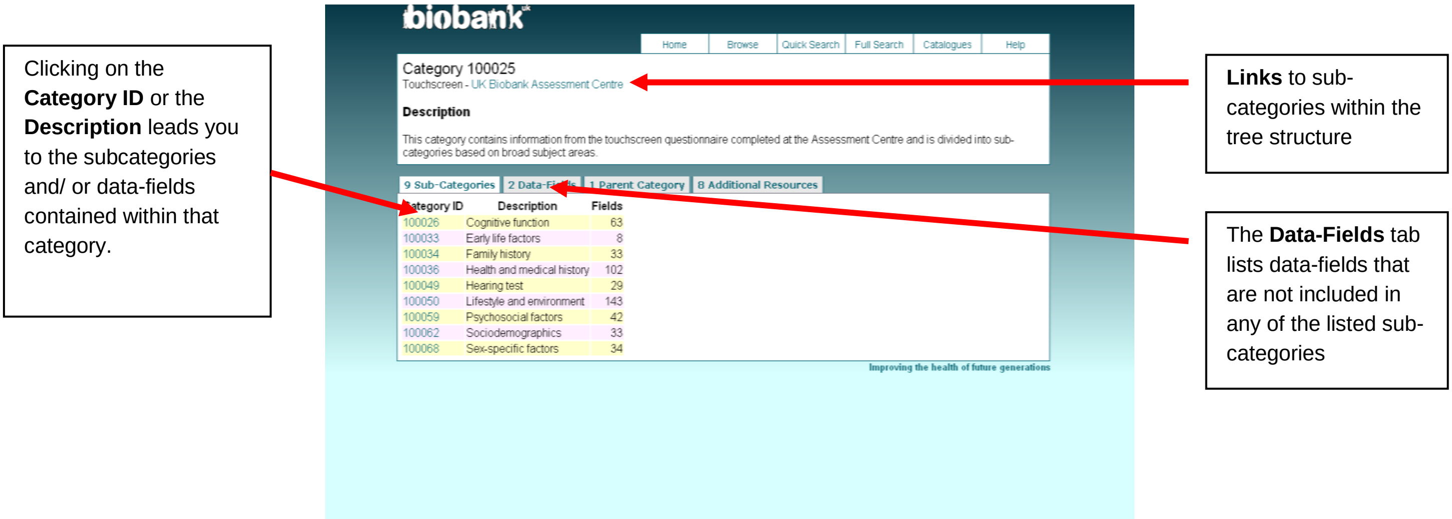
Browse by Primary Category of Origin

Category Description	Fields
Base characteristics	7
UK Biobank Assessment Centre	850
Laboratory biological samples	105
Additional exposure data	0
Health-related outcomes	0

The **Help** button provides more information about items, as listed in the **Glossary** (at the bottom of the Help page)

Clicking on the **‘Show Level’** button is an easy way to jump to a more detailed level

Figure 1. Illustration of the tree structure via ***BROWSE***



Category 100025

Touchscreen - UK Biobank Assessment Centre

Description

This category contains information from the touchscreen questionnaire completed at the Assessment Centre and is divided into sub-categories based on broad subject areas.

Category ID	Description	Fields
100026	Cognitive function	63
100033	Early life factors	8
100034	Family history	33
100036	Health and medical history	102
100049	Hearing test	29
100050	Lifestyle and environment	143
100059	Psychosocial factors	42
100062	Sociodemographics	33
100068	Sex-specific factors	34

Clicking on the **Category ID** or the **Description** leads you to the subcategories and/ or data-fields contained within that category.

Links to sub-categories within the tree structure

The **Data-Fields** tab lists data-fields that are not included in any of the listed sub-categories

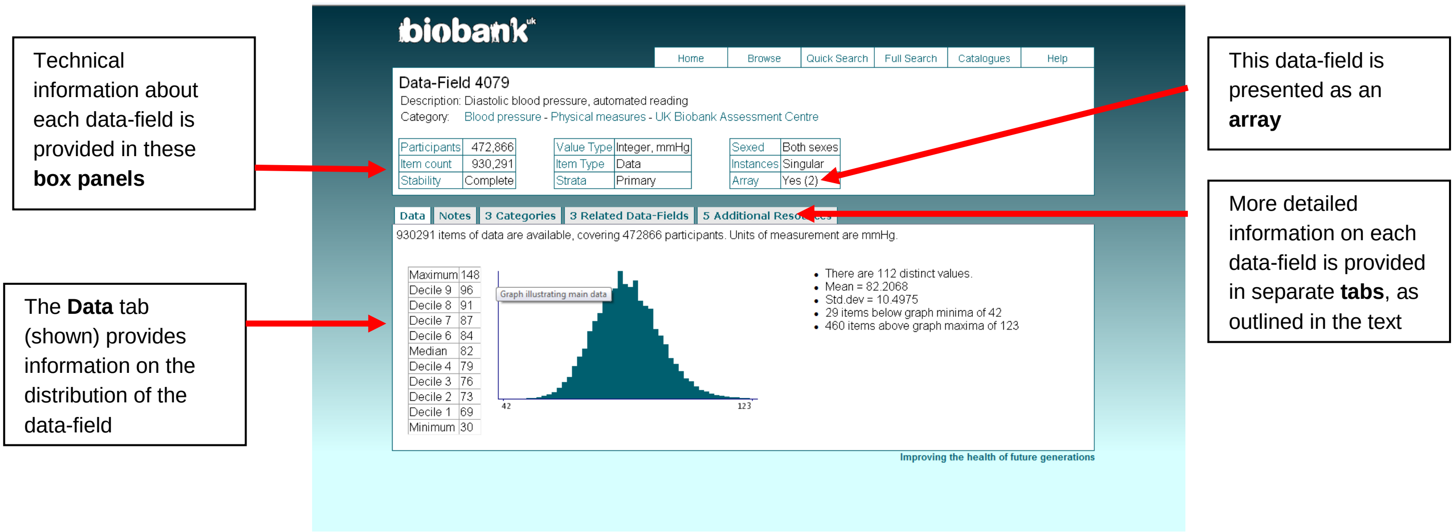
Figure 2. Illustration of sub-categories within the 'Touchscreen' category

Please note that not all data-fields are contained in the most detailed sub-category, since some relate directly to higher (i.e. parent) categories in the tree structure. For example, data-fields related to the time taken to complete the touchscreen questionnaire relate to the touchscreen category as a whole, and are found in the **Data-Fields** tab under the 'Touchscreen' category (see Figure 2).

The tree structure assigns data-fields to one location only, and is not currently cross-referenced. It is therefore important to look in all parts of the tree that might contain data-fields relevant to your research question(s). In general, you should not rely on the **SEARCH** facility to find all fields of relevance for a particular topic.

5 Data-field information

The panel in the top-half of the data-field screen provides a brief description and category location of the data-field within the tree structure (Figure 3). It also includes more detailed technical information about each data-field. This includes information on: the number of participants that have the data item (**Participants**); the number of data items available (**Item Count**); whether the data-field is complete or may change over time (**Stability**); the format and units of the data-field (**Value type**); whether the data-field is a simple data point, relates to an inventory of biological samples, or is a large data object (**Item type**); the likely relevance to researchers of the data-field (**Strata**); whether the data-field is available for both sexes (**Sexed**), how many occasions participants have this measurement performed (**Instances**), and whether there are multiple data items for each instance (**Array**). For example, Figure 3 shows that data on diastolic blood pressure is presented in an array with 2 values per measure (because the measurement was performed twice). Please see the **HELP** page for more details.



Data-Field 4079

Description: Diastolic blood pressure, automated reading

Category: Blood pressure - Physical measures - UK Biobank Assessment Centre

Participants	472,866		Value Type	Integer, mmHg		Sexed	Both sexes
Item count	930,291		Item Type	Data		Instances	Singular
Stability	Complete		Strata	Primary		Array	Yes (2)

930291 items of data are available, covering 472866 participants. Units of measurement are mmHg.

Maximum	148
Decile 9	96
Decile 8	91
Decile 7	87
Decile 6	84
Median	82
Decile 4	79
Decile 3	76
Decile 2	73
Decile 1	69
Minimum	30

- There are 112 distinct values.
- Mean = 82.2068
- Std.dev = 10.4975
- 29 items below graph minima of 42
- 460 items above graph maxima of 123

Technical information about each data-field is provided in these **box panels**

The **Data** tab (shown) provides information on the distribution of the data-field

This data-field is presented as an **array**

More detailed information on each data-field is provided in separate **tabs**, as outlined in the text

Figure 3. Illustration of a data-field

The univariate distribution of each data-field is presented in graphical or tabular format (or both) in the **Data** tab (Figure 3). Distributions of data-fields that are of a sensitive nature (e.g., number of sexual partners) are not shown, although approved researchers can still request such data in their application.

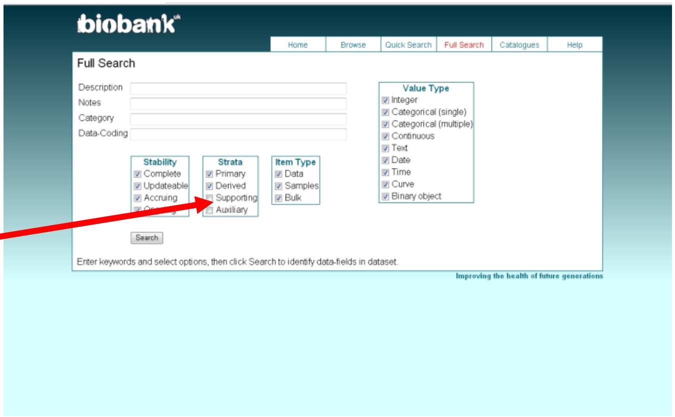
The **Notes** tab includes the full description of the data-field, and for touchscreen questions, provides the exact text of the question that was asked, together with other details.

The **Categories** tab lists the categories and sub-categories of which the data-field is a member. This is also shown horizontally in the category tree, at the top of the page.

The **Related Fields** tab lists other data-fields to which the current data-field is related. For example, the data-field for ‘diastolic blood pressure, automated reading’ (ID: 4079) is related to 3 data-fields: one on diastolic blood pressure from a manual reading, one on systolic blood pressure, and one on pulse taken by the same device (see Figure 3).

The **Additional Resources** tab contains explanatory documentation related to each data-field. This may include screen-shots of the touchscreen questions, details of how each measurement was performed (in downloadable pdf format), photos and video-links.

Some data-fields that are not of primary interest to most researchers may nonetheless be of interest for some research purposes, and these have been classified as supporting or auxillary data-fields (in **Strata**). Examples include the keystroke history of a participant during a touchscreen question, and serial numbers of devices/equipment. Supporting and auxillary data-fields are not searched using **Quick Search,** although you can use the **Full Search** facility to specify the type of data-fields that are included in the search request (see Figure 4). You can also find these data fields using the **BROWSE** function.



Full Search

Description
Notes
Category
Data-Coding

Stability
☑ Complete
☑ Updateable
☑ Accruing
☑ Ongoing

Strata
☑ Primary
☑ Derived
☐ Supporting
☐ Auxiliary

Item Type
☑ Data
☑ Samples
☑ Bulk

Value Type
☑ Integer
☑ Categorical (single)
☑ Categorical (multiple)
☑ Continuous
☑ Text
☑ Date
☑ Time
☑ Curve
☑ Binary object

Enter keywords and select options, then click Search to identify data-fields in dataset.

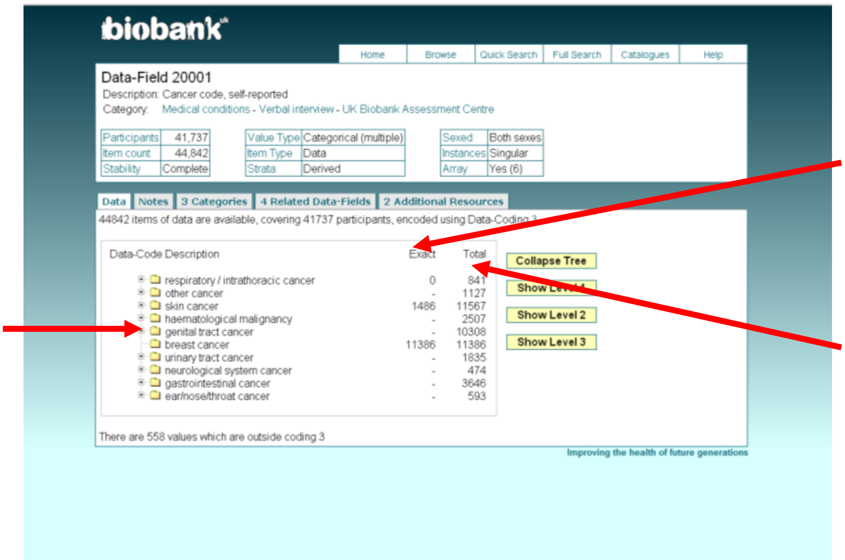
Full Search allows you to specify the criteria of the search. The Quick Search option does not search supporting or auxillary data-fields by default, but these can be specified using Full Search option

Figure 4. Illustration of the Full Search facility

6 Self-reported medical conditions

We advise that you use the **BROWSE** function (rather than **SEARCH**) to find data about a medical condition of interest. Self-reported medical conditions at assessment were indicated on the touchscreen questionnaire, and then confirmed through an interview with a trained member of staff (please see the category description of 'Medical conditions' for more details).

In Figure 5, the '**Exact'** column shows the number of data items listed in each category, as shown. The '**Total'** column shows the number of data items listed in the parent category. For example, there are 11,567 data items for skin cancer, of which 1,486 were coded as simply 'skin cancer'; the remainder could be classified at a more detailed level of the tree as either 'non-melanoma skin cancer' or 'malignant melanoma'.



Data-Field 20001

Description: Cancer code, self-reported

Category: Medical conditions - Verbal interview - UK Biobank Assessment Centre

Participants	41,737
Item count	44,842
Stability	Complete

Value Type	Categorical (multiple)
Item Type	Data
Strata	Derived

Sexed	Both sexes
Instances	Singular
Array	Yes (6)

44842 items of data are available, covering 41737 participants, encoded using Data-Coding 3

Data-Code Description	Exact	Total
⊞ respiratory / intrathoracic cancer	0	841
⊞ other cancer	-	1127
⊞ skin cancer	1486	11567
⊞ haematological malignancy	-	2507
⊞ genital tract cancer	-	10308
breast cancer	11386	11386
⊞ urinary tract cancer	-	1835
⊞ neurological system cancer	-	474
⊞ gastrointestinal cancer	-	3646
⊞ ear/nose/throat cancer	-	593

There are 558 values which are outside coding 3

The **tree** expands when you click on the + box, to give you sub-categories of each condition.

The **'Exact'** column shows the number of data items listed in each category, as shown.

The **'Total'** column shows the number of data items listed in the parent category

Figure 5. Illustration of the coding for medical conditions in the verbal interview

7 Data cleaning

Data from the touchscreen questionnaire have been subject to data checks, as outlined in the explanatory documentation. Data from automatic devices were entered directly into the computer thereby minimizing manual entry of data. Nonetheless, there may be occasions where wrong device numbers were entered, the date-time stamp was incorrect, or there were lapses in calibration. None of the data has undergone further data cleaning at this stage.