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 (with a roughly even number of men and women) recruited between 2006 and 2010 throughout the UK. Showcase (available at 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.ukbiobank.ac.uk .**
- **Take time to familiarise 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.**
- **While it is worth checking what variables are available in the Showcase Resource, there is no requirement to construct a Showcase basket until the Main Application stage.**
- **Note that Showcase is continually under development, as new data on exposure and health outcomes is incorporated into the database. More information on the timelines for future data is availability in the Essential information section of data Showcase.**

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

biobank 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 (a map of the 22 assessment centres is provided in the Essential Information section of the Showcase). 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, 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. Some of these measures were incorporated into the Assessment visit towards the end of the recruitment period and are therefore not available for all 500,000 participants (see the 'Essential Information' section of the Showcase for more information on timelines).

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. 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, family history, qualifications and household income). Pilot data-fields can be identified easily; they include 'pilot' in the data-field name and have a field ID number in the 10000s). 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 lights pattern memory test' on the touchscreen questionnaire and the 'word test' that was performed during the verbal interview stage). A web-based dietary questionnaire was included in the Assessment visit towards the end of the recruitment period, and participants were also invited via e-mail to complete the questionnaire on four further occasions (between 2011 and 2012).

2.2 Data collected at the Repeat Assessment Centre

A repeat assessment of 20,000 participants was carried out between August 2012 and June 2013 at the UK Biobank Co-ordinating Centre, Stockport, UK. Participants who lived within a 35 km radius of the assessment centre were invited to attend an appointment and undergo a repeat assessment of all the baseline measures. These data are now available in data Showcase. For further details; please consult the guide to Repeat Assessment Data (available in the Essential Information, Understanding the UK Biobank study section of the Showcase).

2.3 Enhancement data

In addition to data collected from Assessment Centre visits, the UK Biobank dataset contains additional exposures data and data collected via online questionnaires, including:

- **Online 24-hour dietary recall questionnaire data.** Participants with a known working email address (~320,000 participants) were asked to complete the questionnaire on four separate occasions over an approximate annual period (Feb 2011 – April 2012). The questionnaire was the same as the one administered at the Assessment Centre visit. Please see the Diet by 24-hour recall category (category ID: 100090) for more details.

- **Physical activity data** collected for over 100,000 participants via a wrist-worn accelerometer. Data was collected between June 2013 and January 2016. For further information please see the Physical activity measurement category (category ID: 1008).

- **An online 'Healthy Work' questionnaire** collected data on: occupational history since finishing full time education; respiratory health outcomes and medication for these conditions; and smoking habits. For further details please see the Work environment category (category ID: 123).

- **A web-based questionnaire on cognitive function.** Several of the cognitive function tests administered at the baseline Assessment Centre were re-implemented as web-based questionnaires, along with two additional cognitive function tests. Details of the questionnaire and tests can be found on Data Showcase in the Cognitive function online category (category ID: 116).

- **Web-based questionnaire on mental health.** In 2016, participants with a working email address were invited to complete an on-line mental health self-assessment questionnaire, with the aim being to enrich UK Biobank's phenotyping of mental disorders. The questionnaire covered the most common mental disorders; environmental exposures for mental disorders such as life events, past trauma and substance abuse; happiness and subjective well-being. Information on the questionnaire can be found in the Mental health category (category ID: 136).

2.4 Imaging study data

The UK Biobank Imaging study aims to scan (image) 100,000 UK Biobank participants. At the Imaging Assessment Centre data are collected from Abdominal, Brain and Heart magnetic resonance imaging (MRI) scans, carotid ultrasound scans and 12-lead electrocardiogram (ECG) measurements. The visit also includes whole body DXA imaging, which provides information on body composition, bone size, bone mineral content and bone mineral density.

Imaging data collection began in 2014 at an Imaging Assessment Centre in Cheadle. In 2017 a second Imaging Assessment Centre opened in Newcastle. The data collected thus far are available in Data Showcase in the Imaging category (category ID: 100003).

2.5 Genetics data

Genome-wide genetic data are available for 488,000 UK Biobank participants. Genotype calling was performed on two closely related purpose-designed arrays. There are 805,426 markers in the released genotype data and ~96million genotypes were imputed using computationally efficient methods combined with the Haplotype Reference Consortium and UK10K haplotype resources. Further information about the genetics data available and how to access the data can be found in the Data Showcase Genomics category (category ID: 100314).

2.6 Linked health-related data

Data Showcase also provides information on participants provided through linkage to a range of health-related records, including hospital in-patient data and data obtained from national death and cancer registries. Further details can be found in the Health-related outcomes category (category ID: 100091).

2.7 Future data availability

Please see the 'Essential Information' section of the Showcase for an outline of which measures will be incorporated into the Resource over the next 12-18 months.

3 Finding data in Showcase

RECOMMENDED CATEGORIES: To make it quicker and easier for you to build your dataset, we have created groups of commonly requested data-fields (i.e. variables) to act as a useful starting point so that you don't have to select individual data-fields from each category. These have been loosely grouped into categories such as demographics, cognitive function (that contain the main summary data-fields for all tests), physical measures (that contain the main summary data-fields from all measures), etc. You can then add or remove individual data-fields to or from your pre-populated basket to create a bespoke dataset for your research project. You can find these in the **CATEGORIES** listing in the **CATALOGUES** section (http://biobank.ctsu.ox.ac.uk/crystal/cats.cgi); alternatively, you can link to them from the Essential Information section (Commonly requested categories and fields).

You can find the rest of the data that you need 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: The default search is a text search of the data-field name, its notes and data-codings. Entering a numeric value in the search box will return the data-field with that ID.
The **Data-Field Search** facility also allows you to conduct a search using specific criteria based on the type of data-field (see Section 5 for more details).

Text searches of Data-Codings, Categories, Resources, and Datasets can be conducted by selecting the relevant search type button.

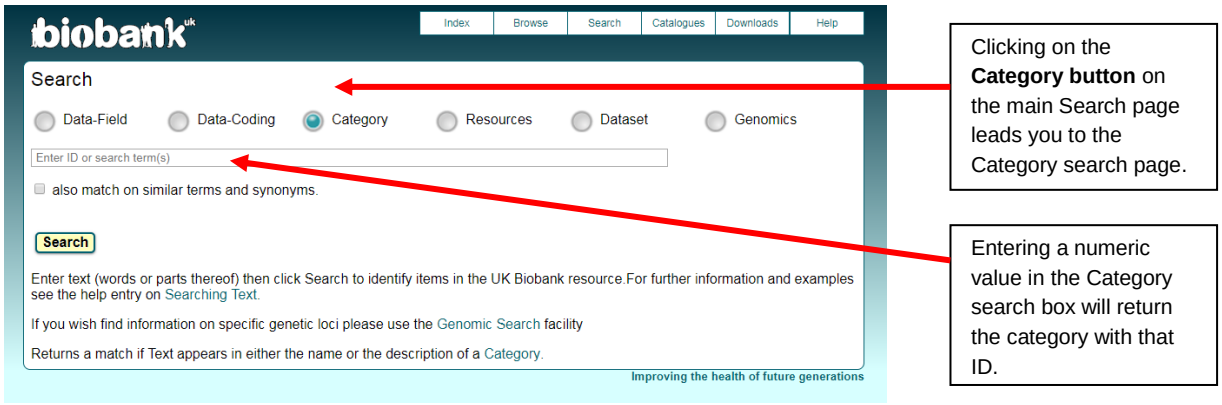


Figure 1. Illustration of the Category Search

Please see the **HELP** tab on the Search page for more details on conducting text searches.

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

The **Genomics search** allows searches of SNPs, the chromosome containing a locus, the position of a DNA segment and the range about those values to include in the search (see Figure 2).

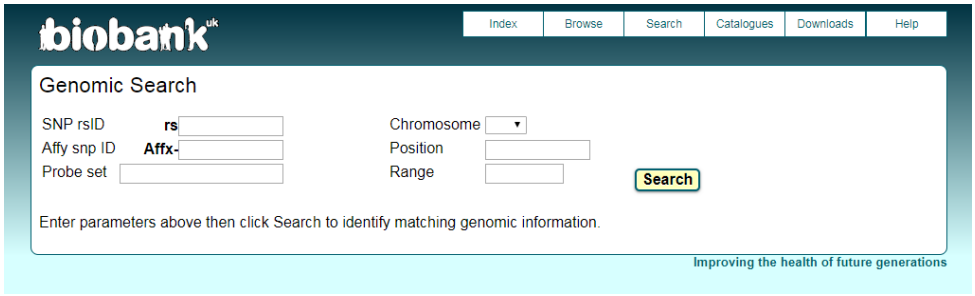


Figure 2. Illustration of the Genomic Search

4 Data categories and sub-categories

Data are organised in a tree structure, accessible via **BROWSE**, with the main categories based on the origin of data collection (Figure 3). These include:

- Population characteristics (some general characteristics of participants known before arrival or ongoing)
- UK Biobank Assessment Centre (data obtained at the Assessment Centre)
- Biological samples
- Genomics (genome-wide genetic data)
- Online follow-up (data obtained using online questionnaires)
- Additional exposures (data collected outside the Assessment Centre)
- Health-related outcomes (data from linkage of participants to health-related records)
- Returned datasets (datasets returned by researchers using the UK Biobank resource).

Please see the **HELP** page on 'Browse' for more details.

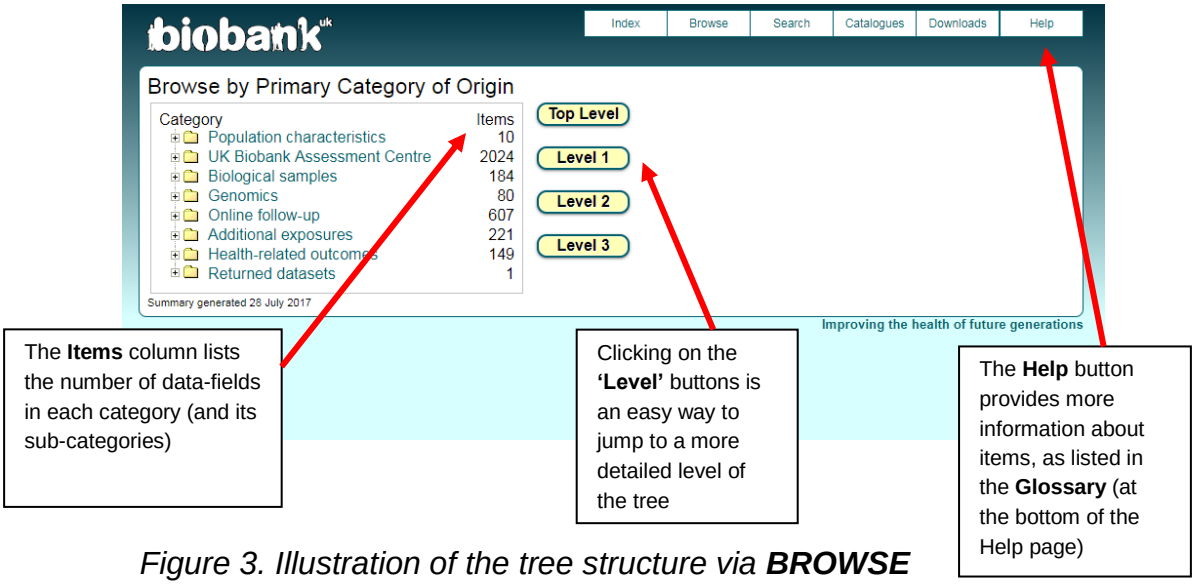


*Figure 3. Illustration of the tree structure via **BROWSE***

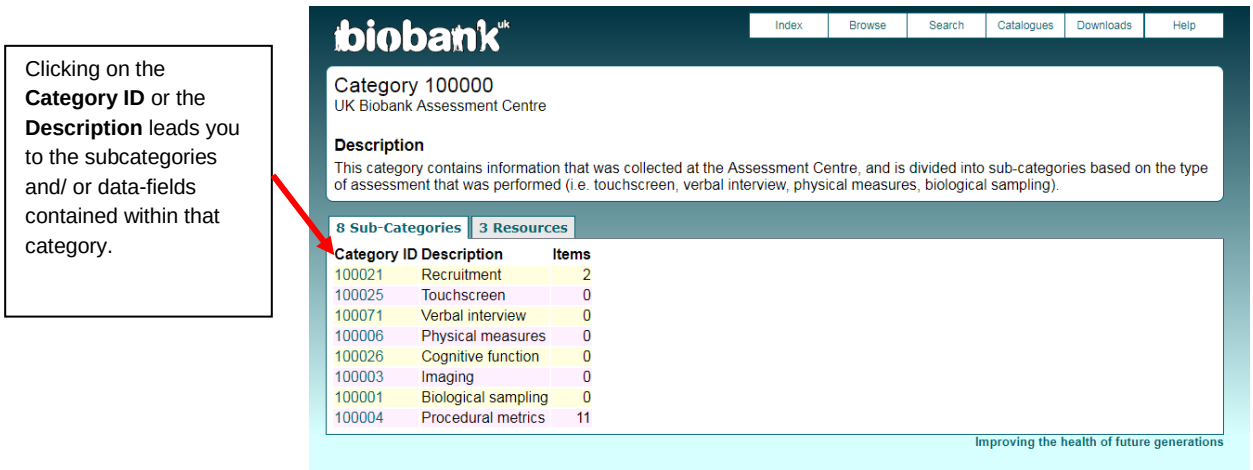


Figure 4. Illustration of sub-categories within the 'Touchscreen' category

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 5). It also includes more detailed technical information about each data-field. This includes information on:

- **Participants**: the number of participants that have the data item
- **Item Count**: the number of data items available
- **Stability**: whether the data-field is complete or changes over time
- **Value type**: the format and units of the data-field
- **Item type**: whether the data-field is a simple data point, relates to an inventory of biological samples, or is a large data object
- **Strata**: the likely relevance to researchers of the data-field
- **Sexed**: whether the data-field is available for both sexes
- **Instances**: how many occasions participants have this measurement performed
- **Array**: whether there are multiple data items for each instance. For example, Figure 5 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.

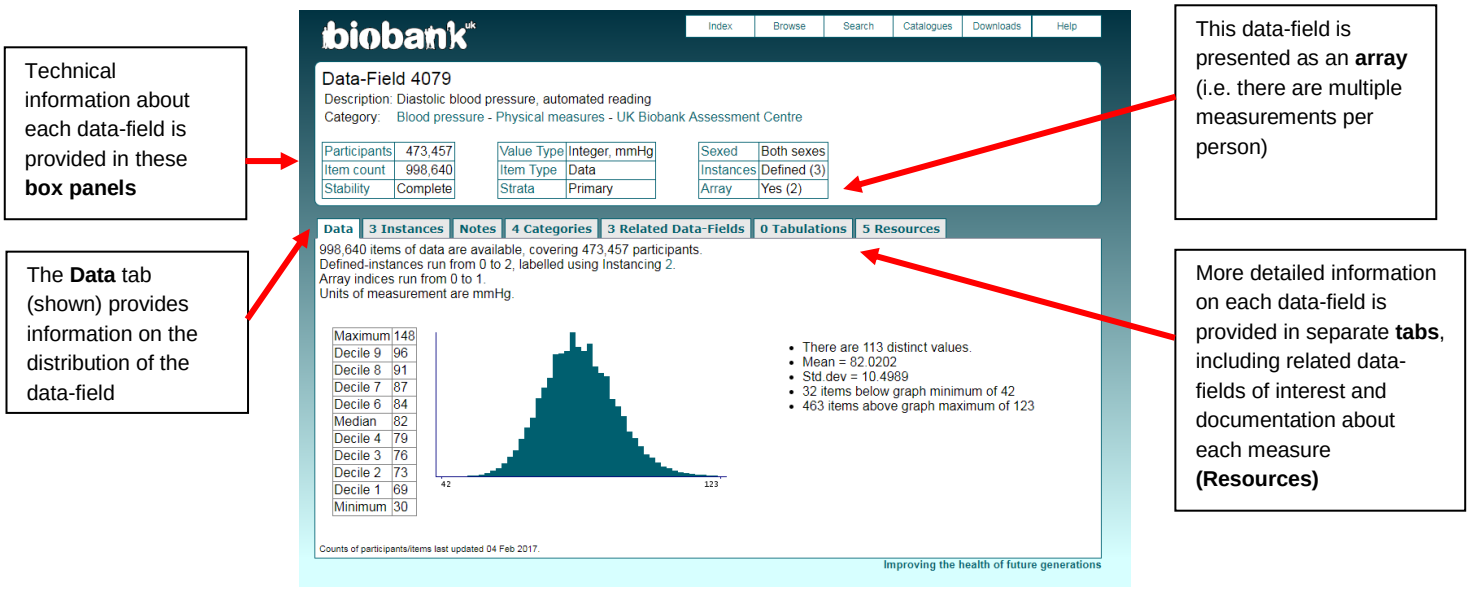


Figure 5. Illustration of a data-field page

The univariate distribution of each data-field is presented in graphical or tabular format (or both) in the **Data** tab (Figure 5). Data are not presented if they are free-text, curve data (i.e. spirometry curves) or bulk data items (i.e. too large to be downloaded: eye images and exercise/ECG results). 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 **Instances** tab provides the univariate distribution of each data-field at each instance (e.g. for Data-Field 4079 the data are presented separately for the initial assessment visit (Instance 0), First repeat assessment visit (Instance 1), and Imaging visit (Instance 2)).

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.

The **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 auxiliary data-fields (in **Strata**). Examples include the keystroke history of a participant during a touchscreen question, and serial numbers of devices/equipment. Supporting and auxiliary data-fields are not searched in the default search**,** but can be included by checking the relevant boxes in the Strata panel. The type of data-fields that are included in the search request can be modified by selecting/deselecting options in the Stability, Strata, Item Type and Value Type panels (see Figure 6). You can also find these data fields using the **BROWSE** function.

Data-Field Search allows you to specify the criteria of the search. The default Search option does not search supporting or auxiliary data-fields by default, but these can be specified using the options in

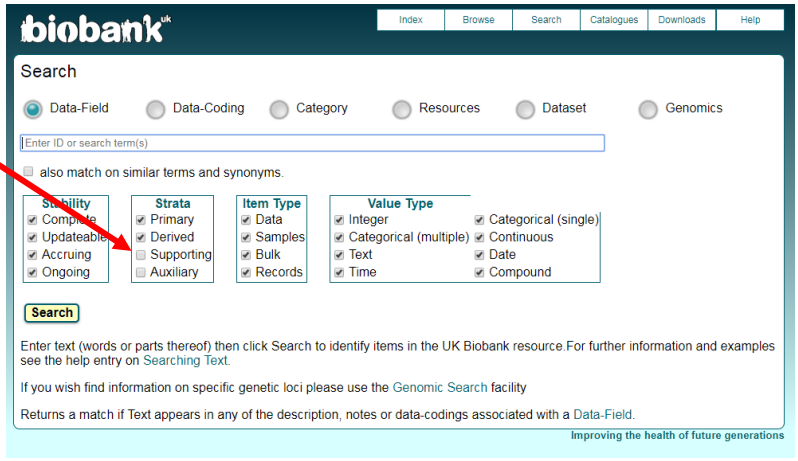
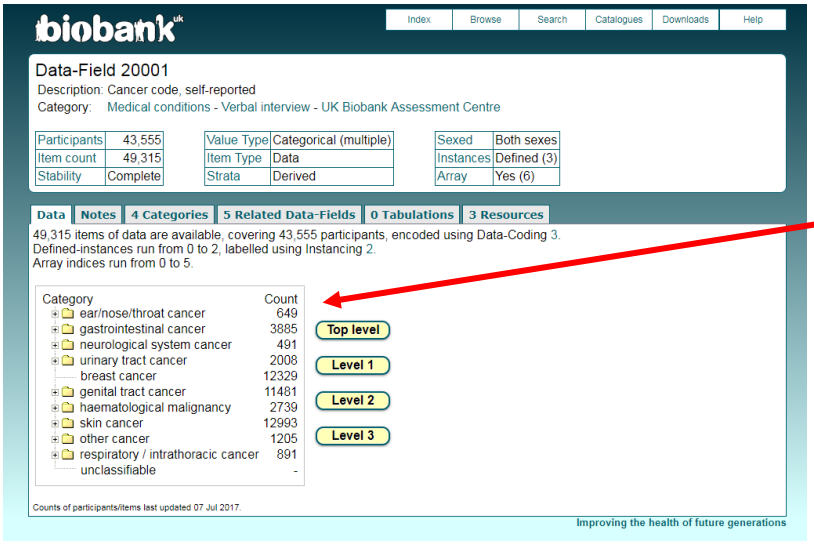


Figure 6. Illustration of the Search facility

6 Self-reported medical conditions

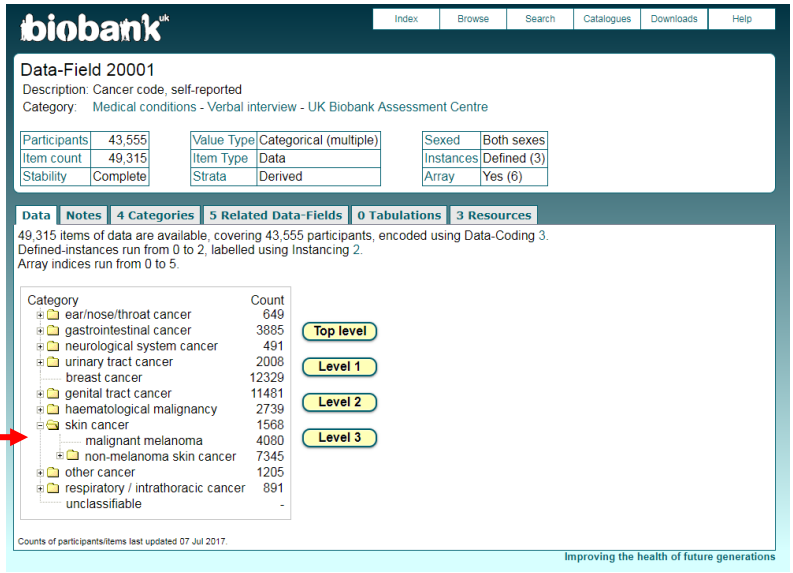
We advise that you use the **BROWSE** function (rather than **SEARCH**) to find data about a medical condition of interest. Self-reported medical conditions at assessment were indicated on the touchscreen questionnaire, and then confirmed through an interview with a trained member of staff (please see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

In Figure 7, the '**COUNT'** column shows the number of data items listed in the category - there are 12,993 data items for parent category 'skin cancer' for example. Clicking on + box reveals more detailed sub-classifications of the condition – e.g. there are 1568, 4080, and 7345 data items for 'skin cancer', 'malignant melanoma' and 'non-melanoma skin cancer' respectively (Figure 8). Please note that the counts displayed are item counts rather than counts of the number of participants who have self-reported the condition.



The **'COUNT'** column shows the number of data items listed in the category.

Figure 7. Illustration of the coding for medical conditions in the verbal interview



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

Figure 8. Illustration of the coding for medical conditions in the verbal interview

7 Health-related Outcomes

UK Biobank acquires information on participants' health outcomes from a variety of different data sources. Death registrations, cancer registrations and hospital episode data are currently being obtained on a regular basis and are being made available to researchers via Data Showcase. These data can be found in the Health-related outcomes category (category ID: 100091).

UK Biobank holds both prospective and retrospective data. Health Outcomes Reports are published periodically, which aim to give researchers an indication of the number of prevalent and incident cases for the most common conditions, by age group, sex and year. For the latest report, please see the Resources tab of the Health-related outcomes category.

8 Algorithmically-defined outcomes

To help researchers include health-related outcomes in their analyses, UK Biobank includes classifications of selected health-related events obtained through algorithmic combinations of information collected at the baseline assessment centre, linked hospital admissions data and death registry information. These algorithmically-defined outcomes are based on algorithms developed by the UK Biobank outcome adjudication group. These data, and more details of the algorithms, are available in the Algorithmically-defined outcomes category (category ID: 42).

9 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 minimising 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. The majority of data that was entered as free-text (e.g., reason for skipping various measures) has been subsequently coded. However, some data-fields (such as serial device IDs for various physical measures) remain as free-text data items owing to the large number of devices that were used.

Data obtained via linkage are also subject to data validation checks and cleaning prior to being made available to researchers. This involves identifying ambiguities in the data, such as invalid clinical classification codes, implausible date values or mismatches of participants' records.

All data imported into the UK Biobank database are validated using the following checks:

- Checks for mismatches
- Checks for data formatting
- Checks against a definitive list of coded values

Values which fail validation are flagged for attention and investigated further.

For further details please see the Validation and cleaning of externally collected data document which can be accessed via the 'Understanding UK Biobank' page.