UK Biobank

Algorithmically-defined health outcomes

Version 1, January 2017

Documentation prepared by:

Christian Schnier (UK Biobank Epidemiologist) & Cathie Sudlow (UK Biobank Chief Scientist),
with input from members of the UK Biobank Follow-up and Outcomes Adjudication Group

This category holds classification of selected health-related events, obtained through algorithmic combinations of coded information from UK Biobank's baseline assessment data collection (which included data from participants on their self-reported medical conditions, operations and medications), along with linked data from hospital admissions (diagnoses and procedures) and death registries. The classification is intended to help researchers to include health-related outcomes in their analyses without having to select lists of diagnostic and/or procedural codes and combine the different data sources themselves. It is based on algorithms developed by the UK Biobank outcome adjudication group, aiming to classify disease outcomes with high positive predictive value (i.e. a high probability that people classified as being positive for a health-related event have indeed experienced that event). Where possible, we provide the best available information on the estimated positive predictive value of each source or combination of sources so that researchers can use this information in their analyses.

Process

Information used for the classification is taken from the baseline assessment touchscreen questionnaire and/or nurse-led interview and from linked national health-related datasets. Data from primary care linkages is not currently included but will be incorporated into future versions of the algorithms. To classify individuals, the relevant datasets were interrogated for disease-related codes. The algorithms used to select and combine disease-related codes from the different sources are provided separately for each specific condition or group of related conditions.

Information on disease classification has been incorporated as data fields that can be added to UK Biobank datasets. These fields hold, for any of the selected health-related events, information on the date of event and the source. The table below gives some examples:

ID	**Any stroke**		**Subarachnoid haemorrhage**		**...**	**Any MI**		**...**
	Date	**Source**	**Date**	**Source**	**...**	**Date**	**Source**	**...**
007	01.01.2015	Self-Reported only						
009						01.03.2015	Hospital admission	
...	...	...						

Daterefers to the earliest date of the health related event found in any of the combined datasets; and **Source** refers to the dataset in which the event was identified. These are defined as follows: **Source** is classified as '**Self-report only**' for participants who indicated in the nurse-led interview at their baseline assessment that they had experienced the event and who had no relevant hospital admission diagnosis or procedure code prior to their date of recruitment into the UK Biobank cohort. The **Date** for '**Self-report only**' events is derived from information

obtained at the nurse led interview conducted at the baseline assessment[1]. **Source** is classified as '**Death only**' for participants with a relevant code in the death registration records with no baseline self-report of the event and no relevant hospital admission diagnosis or procedure code. The **Date** for '**Death only**' events is the date of death. **Source** is classified as '**Hospital admission**' for participants with a relevant diagnosis or procedure code in the hospital admission data (irrespective of whether they also self-reported an event at their baseline assessment in the nurse interview or had a relevant code in the death registry records). The **Date** for **'Hospital admission'** events is the earliest date of a relevant event within the linked hospital admissions data.

Using the data fields

The algorithms are designed to enable the selection of cases of disease for a range of different research study designs. The prospective design of UK Biobank makes it particularly suitable for studies involving incident cases (i.e. those first diagnosed or detected with a condition after recruitment to the study), but the algorithms identify disease cases diagnosed both before and after recruitment to UK Biobank. Researchers are advised to merge the algorithmically derived outcome data fields with information on date of the participant's baseline recruitment (Date of attending assessment centre) to enable outcomes to be further classified into 'prevalent' (for cases that occurred before recruitment) and 'incident' (for cases that occurred after recruitment). According to current definitions, self-reported events can only be 'prevalent', since information from repeat assessment nurse-led interviews are not included in the algorithms at present, while mortality records will only inform on 'incident' events.

UK Biobank's current algorithms inform only on the earliest health-related event of any particular type. To analyse recurrent events, researchers are advised to download all health-related information and develop their own algorithm. (NB some of the codes in the code list used in this classification might not be indicated for use in analysis of recurrent events). Please be aware that the total number of people with a health-related outcome with source 'Hospital admission' does not reflect the total number of participants who have had a health-related hospital admission. Similarly, the total number of people with a health-related outcome with source 'Death only' does not reflect the total number of participants who have died of that condition; rather, it is the number of people who died with the code(s) on their death record, who did not also have either a hospital admission or self-report event with the relevant code(s). For analysis stratified by source or for summary statistics by source, researchers are advised to download all health-related information and use the codes suggested in the algorithms for the condition(s) of interest.

[1] as either the age of the participant at the time of the event or the year of the event, converted into an exact date using the procedure described under the Notes tab.

Key points to note

- Great effort has been made to provide the optimal algorithm for the majority of potential research studies. However, different research studies might benefit from using alternative algorithms.
- Estimates of disease frequency in the UK Biobank cohort are not representative of the general (British) population
- The algorithms and associated data fields will be updated as additional linked data (especially linked primary care data) are incorporated into the UK Biobank dataset, and with updated information the classification of an individual might change.
- The Outcome Adjudication Group is working on a large range of health related outcomes. Once new algorithms have been developed, new disease classifications will be added to this category.
- Different national data sources were used to classify the health status. Each source provides information for a different range of dates.