UK Biobank

Health Outcomes Overview

Version 3.0

http://www.ukbiobank.ac.uk/

September 2024

Contents

1. Data sources

Information about health outcomes in UK Biobank comes from a variety of different sources. These include:

(1) self-reported data collected at the UK Biobank assessment centre visits;
(2) linkage to electronic health records, such as hospital inpatient records, primary care (GP) records, and death and cancer registrations;
(3) self-reported data collected from online questionnaires for conditions that are often not captured in health records, such as mental health, gastrointestinal conditions, and pain.

2. Types of data available

UK Biobank incorporates linked health data into the resource in a variety of ways including:

Type	Overview	Access method[1]
Standard data-fields	• Cancer registration data • Death registration data	Released as data fields in the main dataset (i.e. alongside the majority of the phenotypic data-fields)
Summary hospital inpatient fields	• Shows all the distinct diagnoses and procedures each participant has recorded throughout their hospital admissions, and the dates these were first recorded in the inpatient data.	
Algorithmically-defined outcomes	• Combines data from hospital inpatient records, conditions self-reported[2] at the baseline assessment visit, and death registrations. • Provides the date the condition was first recorded in any source. • Uses code lists informed by the literature and developed by clinicians. • Currently available for a subset of conditions.	
First occurrence fields	• Maps diagnostic codes to broad health outcomes defined by 3-character ICD10 codes. • Provides the date each health outcome was first recorded in any source. • Provides the source of the earliest record of the health outcome and whether it was also recorded in any other source. • Derived from hospital inpatient records, primary care (GP) records, and conditions self-reported[2] at assessment centre visits.	
Online follow-up questionnaires	• Includes self-reported diagnoses, procedures and/or medications for conditions relevant to the specific questionnaire.	
Record-level data	• Contains full hospital inpatient records (not just summary information), death registrations & primary care (GP) records.	Data portal

[1] You can find more information about how data items are served by referring to the Data Access Guide (in particular sections 2.1 and 4.1 for Main Dataset and Data Portal explanations, respectively).

[2] Category 100074 contains the main data-fields related to self-reported medical conditions, however for certain conditions (e.g. eye, mental health) data-fields in other categories of Data Showcase might be relevant. Similarly, additional data on some conditions has been collected via online questionnaires. These are discussed later in this document.

3. Coverage of linked health data

For information on the time periods and participant coverage available for each type of linked health data by data provider, please see the Data Providers & Dates page on Essential Information.

The time periods covered by linked datasets vary by the type of data and the source. For example, whilst Scottish hospital inpatient data is available from around 1981, English hospital inpatient data is available from only around 1997. Furthermore, completeness and accuracy in health linkage data cannot be assumed and is expected to differ between systems and over time.

Linked health data can have other limitations such as coverage or availability. For example, the primary care (GP) data (Category 3000) is available for about 45% of the cohort and is available for all research purposes; whereas the COVID-19 vaccination data (Category 999, fields 32040 and 32041) covers the majority of the cohort but can be used only for COVID-19-related research purposes.

For detailed information on a specific dataset's limitations, please refer to the Resources tab of Showcase for that category of data. The appendices provide links to each data category for further reading and include a brief summary of the caveats associated with each linked data item.

4. Choosing health outcome related data-fields

To help select which health outcome data might be best suited to your research project, researchers may wish to consider the following questions, and then refer to the tables in the appendix. These tables show the location of relevant data-fields on Showcase, caveats associated with the data, and links to further resources for diagnoses, procedures/operations and medications respectively.

Possible considerations when selecting health outcome data:

- Which sources of diagnostic data (HES, primary care, self-report) are appropriate for the conditions your research study covers? Bear in mind that some conditions are unlikely to require hospital inpatient treatment, so might be missed if you only use hospital inpatient data.
- Data from some sources are only available for a subset of the cohort (for example, primary care data is currently available for ~45% of the cohort, with a censoring date of approximately 2016/2017).
- How critical is the date of diagnosis information for your study? If the precise dates are crucial you might not want to use the summary fields: these use the raw 'uncleaned' data, which may have dates that are inaccurately recorded. You may wish to explore the record-level data so you can make your own decisions about data-cleaning.
- Does your research require information on all occasions when the diagnosis was recorded, or just the first? If the former, then you will need to access the record-level data.
- Are you interested in the primary diagnosis only (the main reason for the hospital admission), or primary and secondary diagnoses (both the main reason and underlying conditions)?
- For some conditions, we have generated algorithmically determined outcomes, where the code lists have been developed in collaboration with clinicians.
- Do you want to adopt a 'belt and braces approach' where you look across all data sources that mention diagnoses? For example, we normally recommend that researchers identify participants with cancer using the cancer register as the 'gold standard'. But if you want to be as certain as possible that the participants in your cohort have <u>not</u> been diagnosed with cancer you might want to use the cancer registry + inpatient hospital data + self-report + GP data.

Appendix A: Information on diagnoses

Data Source		Category ID	Description	Caveats	Links to more information
Self-report at assessment centre		100074	Medical conditions self-reported during the verbal interview.	• Only contains conditions the participant reports, so might not include certain conditions, or more minor conditions, or those diagnosed many years before	Resource 100235 — The verbal interview within ACE centres
		100044[3]	Medical conditions self-reported via the touchscreen	• Might not be as reliable or complete as the medical condition data collected in the verbal interview with a nurse	Resource 113241 — Touchscreen questionnaire ordering, validation and dependencies
Hospital inpatient data	Summary data-fields	2002	Lists all the diagnoses recorded in the record-level data, and the date each code was first recorded[4]. Separate data- fields are available for diagnoses recorded using ICD10 and ICD9 (for older hospital records), and also for primary diagnoses, and diagnoses recorded as either a primary or secondary diagnosis.	• Derived from raw data that may contain inconsistent dates etc. • No psychiatric or maternity data are yet available for admissions to Scottish hospitals	Resource 141140 — Inpatient data Dictionary Resource 138483 — Hospital Inpatient data
	Record-level data-fields	2006	These data-fields grant access to the record-level data.	• Raw data that has only been subject to minimal data cleaning by UK Biobank	
Primary Care	Record-level data-fields	3001	These data-fields grant access to the record-level data.	• Currently available for about 45% of the cohort • Data is not currently accruing • There has been no detailed data cleaning	Resource 591 — Primary Care data
Cancer Register		100092	Information on the type of cancer (ICD-10, ICD-9), histology and behaviour of the tumour are currently available	• There has been no detailed data cleaning	Resource 115558 — Linkage from National Cancer Registries

[3] We recommend data-fields in category 100074 are used in preference to those in 100044.

[4] The inpatient hospital data does not record diagnosis date, but the episode start date (or admission date if the former is missing) will provide a proxy for this.

Data Source	Category ID	Description	Caveats	Links to more information
Death Register	100093	• Information on the underlying (primary) and contributory (secondary) causes of death obtained from linkage to national death registries. • Coded with ICD-10	• There has been no detailed data cleaning	Resource 115559 — Linkage from National Death Registries
COVID-19 test data (linkage)	999	Records of COVID-19 test results from linkage to national records	• Dataset evolved as testing scaled and more community testing was introduced • Reporting of negative results may be inconsistent • There has been no detailed data cleaning	COVID-19 test result Data Dictionary
COVID-19 antibody test data (self-administered test study)	998	Information on antibody test results for COVID-19	• Covers about 40% of the cohort • Tests were self-administered • Data was collected across two phases	Resource 4500 - COVID-19 antibody study Phase 1 Resource 4501 - COVID-19 antibody study Phase 2
Algorithmically-defined outcomes	42	• Combines data from inpatient hospital records, conditions self-reported at an assessment centre visit and death registrations (future developments will also incorporate primary care). • Provides the date the condition was first diagnosed • Uses code lists informed by the literature and developed by clinicians	• Only available for certain conditions/events: Asthma, COPD, Dementia, End stage renal disease, Motor neurone disease, Myocardial infarction, Parkinson's disease, Stroke	Resource 460 - Algorithmically defined outcomes[5] Resource 8319 - Algorithmic ESRD definition
First occurrences	1712	Identifies diagnostic codes and the date it was first recorded in inpatient hospital records, primary care or conditions self-reported at the assessment centre, mapped to broad health outcomes defined by 3-character ICD10 codes.	• The mapping between the source Read2/Read3/ICD9 code to 3-character ICD10 is not complete. • The algorithm uses the raw, uncleaned data	Resource 593 - First Occurrences of Health Outcomes Defined by 3-character ICD10 code
Self-reported eye problems	100041	As part of the touchscreen assessment participants were asked to indicate whether a doctor had told them they had any eye conditions (e.g. myopia, hypermetropia, astigmatism, strabismus, glaucoma, cataract, macular degeneration)	• Might not be as reliable or complete as the medical condition data collected in the verbal interview with a nurse	Resource 100584 - Screenshot from touchscreen questionnaire used to capture field 6148

[5] This document covers definitions for most of the conditions listed since they are similar to each other. ESRD's definition is sufficiently different to warrant its own definition document, also linked below this item.

Data Source	Category ID	Description	Caveats	Links to more information
Self-reported hearing problems	100043	As part of the touchscreen assessment participants were asked to indicate if they had difficulty with hearing or were completely deaf.		
Self-reported dental conditions	100046	Information on mouth/teeth problems collected via the touchscreen at the assessment centre visit.		
Self-reported pain	100048	At the touchscreen assessment participants were asked various questions about pain and whether they had experienced headaches.		
Self-reported mental health	100060	Information on mental health collected via the touchscreen at the assessment centre visit.		
Online follow-up - mental health self-assessment questionnaire	136	Included questions asking participants whether they had ever been diagnosed with mental health problems by a professional (see data-field 20544).	• Data available for only participants who completed the questionnaire	Resource 22 - Mental health web-based questionnaire
Online follow-up - work environment	132	Included questions about doctor diagnosed respiratory conditions (e.g. COPD, asthma, lung cancer)	• Data available for only participants who completed the questionnaire	Resource 8349 - Occupational history web-based questionnaire
Online follow-up - digestive health	153	Included questions about irritable bowel syndrome and coeliac disease	• Data available for only participants who completed the questionnaire	Resource 595 - Digestive health web-based questionnaire
Online follow-up – mental well-being	1500	Included questions about mental health conditions diagnosed by a professional, and about COVID-19 diagnoses	• Data available for only participants who completed the questionnaire	Resource 2800 - Online mental well-being questionnaire
Online follow-up – experience of pain	154	Included questions about various medical conditions that cause or exacerbate pain	• Data available for only participants who completed the questionnaire	Resource 2718 - Online experience of pain questionnaire

Appendix B: Information on procedures/operations

Data Source		Category ID	Description	Caveats	Links to more information
Self-report at assessment centre		100076	Operations self-reported during the verbal interview	• Only contains operations the participant reports, so might not include all operations, or more minor operations, or those that took place many years before	Resource 100235 - The verbal interview within ACE centres
Inpatient hospital data	Summary data-fields	2005	Lists all the procedures recorded in the record-level data, and the date each code was first recorded[6]. Separate data-fields are available for diagnoses recorded using OPCS4 and OPCS3 (for older hospital records), and also for primary procedures, and procedures recorded as either a primary or secondary procedure.	• Derived from raw data that may contain inconsistent dates etc.	Resource 141140 — Inpatient data Dictionary Resource 138483 — Hospital Inpatient data
	Record-level data-fields	2006	More detailed information on the hospital episodes in which procedures took place.	• Raw data that has only been subject to minimal data cleaning by UK Biobank	
Primary care data	Record- level data- fields	3001	Some procedures taking place in hospitals might be captured within the coded GP records via Read codes. These will be found in the clinical table (access via data-field 42040)	• Only ~45% of the cohort have linked GP records • Only procedures that the GP or other primary care staff have recorded via Read codes on the primary care record will be captured • Data is not currently accruing	Resource 591 - Primary Care data
COVID-19 vaccination records		999	Lists all administered COVID-19 vaccines for English participants	• For COVID-19 research only • English data only • Administered vaccines only	Resource 2910 – COVID-19 vaccination record data
COVID-19 vaccination data from antibody study		998	Includes self-reported first and second COVID vaccination dates	• Covers about 40% of the cohort • Self-reported dates, likely less accurate than Category 999 data	Resource 4500 - COVID-19 antibody study Phase 1 Resource 4501 - COVID-19 antibody study Phase 2

[6] The inpatient hospital data records operation dates for some participants, but for some records this is missing. The episode start date (or admission date if the former is missing) will provide a proxy for this.

Appendix C: Information on medications

Data Source		Category ID	Description	Caveats	Links to more information
Self-report at assessment centre		100075	Information on regular prescription medication self- reported at the assessment centre during the verbal interview with a nurse.	• Only contains prescription medications that the participant reports taking regularly so does not include short-term medications or prescribed medication that is not taken.	Resource 100235 - The verbal interview within ACE centres
		100045	Information on medications reported in the touchscreen.	• Data from category 100075 is more detailed and might be more reliable as it was collected during the verbal interview with a nurse.	Resource 113241 - Touchscreen questionnaire ordering, validation and dependencies
Primary care data	Record-level data-fields	3001	These will be found in the scripts table (access via data-field 42039).	• Only ~45% of the cohort have linked GP records • Only medications that the GP or other primary care staff have recorded via Read codes on the primary care record will be captured • Data is not currently accruing	Resource 591 - Primary Care data
Online follow-up - work environment questionnaire		132	Some information on recent medication for respiratory conditions.	• Data available for only participants who completed the questionnaire	Resource 8349 - Occupational history web-based questionnaire
Online follow-up - mental well-being questionnaire		1500	Self-reported medication for mental health conditions (see data-fields 29038, 29039)	• Data available for only participants who completed the questionnaire	Resource 2800 – Online mental well-being questionnaire
Online follow-up - sleep questionnaire		208	Self-reported over-the-counter and prescription sleep medication (see data-fields 30463, 30464)	• Data available for only participants who completed the questionnaire	Resource 2278 – Online sleep questionnaire

Data Source		Category ID	Description	Caveats	Links to more information
Inpatient hospital data	Summary data-fields	2005	While the inpatient hospital data doesn't include variables related to medications directly, some of the procedure codes relate to the administration of medication in hospital (e.g. OPCS4 code X81 'High cost gastrointestinal drugs'). See the OPCS code lists for details. Lists all the procedures recorded in the record-level data, and the date each code was first recorded[7]. Separate data-fields are available for diagnoses recorded using OPC4 and OPCS3 (for older hospital records), and also for primary procedures, and procedures recorded as either a primary or secondary procedure.	• Fields do not record all medications administered in hospital • Derived from raw data that may contain inconsistent dates etc.	Resource 141140 — Inpatient data Dictionary Resource 138483 — Hospital Inpatient data
	Record-level data-fields	2006	More detailed information on the hospital episodes in which procedures at which drugs where administered took place.	• Raw data that has been subject to only minimal data cleaning by UK Biobank	

[7] The inpatient hospital data does not record an exact prescription or medication administration date. The episode start date (or admission date if the former is missing) will provide a proxy for this.