UK Biobank

Cancer data: linkage from national cancer registries

Version 2.0

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

July 2023

This resource details the data collected via linkage from national cancer registries for UK Biobank participants.

Contents

1. Summary of changes in this version

This document has been extensively revised to include information on recent changes to the cancer registries that we link to, expanded information on censoring and instancing, and links to other relevant documents.

2. Understanding the dataset

UK Biobank provides linked health data for all participants, including deaths, cancers, and hospital inpatient records, which are gathered from several separate national data providers. This document covers information specific to cancer registry linkage, and the structure of the cancer data published.

Over the lifetime of the UK Biobank, several different registries have been used, reflecting changes in record-keeping bodies active in different parts of the UK. A list of registries used, along with the dates of linkage, is provided in Appendix A.

Cancer registries acquire information on cancer diagnoses from a variety of sources including hospitals, cancer centres and treatment centres, hospices and nursing homes, private hospitals, cancer screening programmes, other cancer registers, general practices, death certificates, Hospital Episode Statistics (HES), and Cancer Waiting Time (CWT) data. In many instances, more than one source of information is available to cancer registries from a single organisation, for example hospital patient information systems, pathology laboratories, medical records departments and radiotherapy databases.

UK Biobank receives details of cancer registrations both prior to the inception of study (i.e. cancers dating back to the early 1970s when cancer registries were first established) and following recruitment into the study. Please see the accompanying Cancer Summary report, available under Understanding UK Biobank in Essential Information, which gives information on both prevalent and incident cases, for more details about the numbers of cancers currently in the database.

Aside from linkage data on cancer registrations, UK Biobank also publishes self-reported data on cancer from verbal interviews at assessment centres. This data includes cancer codes and approximate date and participant age at diagnosis, and is available in Category 100074. Self-reported data is not discussed further in this document.

2.1. Data coverage

Data from the registries is sent to UK Biobank periodically, and then incorporated into the resource. The period between updates varies between providers but is typically 6 or 12 months. This, along with delays occurring before the record appears in the relevant cancer

registry (usually at least 6 months and sometimes much longer) and processing time, means that data is not typically available to researchers until at least a year after diagnosis, and in some cases much longer.

A suggested censoring date is published by the UK Biobank, separately for England, Wales and Scotland (as these data come from different registries). For more information on censoring dates including the rule used for estimation, please see the Data providers and dates page on Essential Information.

3. Data available

The data presented in Showcase (category 100092) and available to researchers comprises:

Field id	Field name	Comments	Encoding
40005	Date of cancer diagnosis	From cancer registry	
40006	Type of cancer: ICD-10	From cancer registry	Data-coding 19
40013	Type of cancer: ICD-9	From cancer registry	Data-coding 87
40011	Histology of cancer tumour	From cancer registry	WHO download*
40012	Behaviour of cancer tumour	From cancer registry	
40008	Age at cancer diagnosis	Calculated from date of cancer diagnosis and date of birth	
40009	Reported occurrences of cancer	Calculated: simple count of the number of cancer records for this participant	
40021	Cancer record origin	Broad grouping of registry providers into England & Wales, Scotland, or NCIN (England)	Data-coding 1970
40019	Cancer record format	File format of received data. More detailed indicator of the data source, differentiating between different data formats from the same provider.	Data-coding 262

** Data Fields 40011 and 40012 should be interpreted in combination – see Section 4.2 for more information*

3.1. Instancing

For participants with more than one recorded cancer diagnosis, the set of data relevant to each cancer record is linked via a system of 'instancing', where each occurrence of cancer is presented as a separate 'instance'.

Each instance will appear in its own set of linked columns in the main dataset. For example the date recorded in field 40005_4_0 (date field, instance 4, array index 0) corresponds to the ICD code recorded in field 40006_4_0 (ICD10 field, instance 4, array index 0), and these fields will be blank for participants with three or fewer cancer diagnosis records. Please see Section 5.1 for information on data quality and limitations on instancing.

4. Clinical classifications

4.1. Cancer site

The type of cancer is coded according to the International Classification of Diseases (ICD) from the World Health Organisation (WHO), which provides a system of diagnostic codes for classifying diseases and is revised periodically to account for newly emerging conditions.

As the cancer registry records go back to the 1970s, the type of cancer is coded according to the version of ICD coding that was relevant for that time period. A small number of recorded cancers were diagnosed before 1979, when the eighth revision ICD (ICD-8) codes were in use. All of these particular ICD-8 codes were retained when using ICD-9 format; hence, we have grouped these cancers into the ICD-9 tree structure. The ninth revision (ICD-9) was implemented in 1979 and was replaced by the 10th revision (ICD-10) in 2000-2001.

The ICD-codes are presented in a tree-like structure, grouped according to ICD-10 chapter order (rather than by the diagnosis name). The authoritative source for details on this coding system is the WHO.

4.2. Histology and behaviour of cancer tumour

The morphology of tumours is presented via five-digit codes in the International Classification of Diseases for Oncology, 3rd Revision (ICD-O-3), ranging from M-8000/0 to M-9989/3. The first four digits (after the M) code the histology and the fifth digit codes the behaviour. These codes are represented in the UK Biobank as separate variables (e.g. morphology code "M-8120/3" has its histology component stored in Field 40011 as "8120" and its behaviour stored in Field 40012 as "3").

Please note that ICD-O morphology codes have undergone multiple revisions over time, and it is often impossible to determine which version of the system has been used. For more information about the ICD-O-3 system, please see the WHO documentation.

Previously Showcase encodings were supplied for each field separately. However, the ICD-O system only assigns meanings for a combined morphology code (e.g. "8120/3"). For example, the full morphology code in ICD-O-3 for histology code "8120" offers four different meanings depending on the behaviour code attached:

8120/0: "Transitional cell papilloma, benign"
8120/1: "Urothelial papilloma, NOS"
8120/2: "Transitional cell carcinoma in situ"
8120/3: "Transitional cell carcinoma, NOS"

The separate encodings have thus been retired and researchers should be careful to always refer to the ICD-O-3 meaning for the combined morphology code. A text file of the most up-to-date ICD-O-3 classification is available from the WHO Classifications Download Area (requires an account). ICD-O-3 morphology codes are also detailed in Resource 126048.

ICD-O-3 directs coders to use the appropriate behaviour code even where the resulting combined term does not appear in the ICD-O guide; for example, a code of "9000/2" could be used for "Brenner tumour in situ," if such an entity were to be identified, even though this does not appear in the code lists.

In some cases, errors in either the histology or behaviour codes have been noted during processing. As these are best interpreted as a single morphology code (see above), behaviour codes are redacted where the accompanying histology code was missing or invalid, and vice versa.

5. Data quality

When using the cancer registry dataset, it is worth considering that the following can change over time:

- The versions of the clinical classifications used (see Section 4)
- The range of values and/or admissible codes within data-fields, due to improvements in data cleaning and/or availability of more detailed codes
- Certain data-fields may change, be added or become obsolete over the years.

Particular data quality issues are noted in the following subsections.

5.1. Amalgamation of multiple sources

Cancer registry information is amalgamated from multiple sources, both within UK Biobank and within the registries. Data structures, completeness and/or quality may differ slightly between the sources.

Caution must be exercised in particular when interpreting multiple cancer records for the same participant. In some cases multiple records will reflect multiple diagnoses, but we cannot exclude the possibility of this sometimes including (pseudo-) duplicates between the different providers. While exact duplicates are excluded, two records may actually be duplicates with a minor variation in date, a correction to the ICD code, or other changes made by the registry. A diagnosis may also have been recorded with an ICD-9 code from one provider and then converted to ICD-10 in another, resulting in duplication.

Researchers are therefore advised to treat the merged data with care. In particular, the record counts in Field 40009 should not be treated as the number of distinct cancer diagnoses per patients, since they reflect simply the number of records (which may have pseudo-duplicates as mentioned above).

5.2. Break in coverage for Scottish participants

Due to changes at the data provider, we were unable to link to cancer registry data for Scottish participants for a period starting in 2015. Once linkage was re-established it was clear that some records were missing. In 2022 this gap was filled as far as possible by extracting the relevant fields from a separate registry held by Public Health Scotland, which holds more detailed (enhanced) cancer records. We combined records received in extracts from their database received in 2018 and 2020.

There was some risk of duplication of records since the missing records do not cover a discrete time period. The approach taken was to include every record provided in the enhanced cancer registry unless:

- A record exists in the standard data for this participant with the same date and the same ICD10 code OR
- A record exists in the standard data for this participant with the same date and ANY ICD9 code (all enhanced records use ICD10 codes) OR
- Participant has withdrawn consent

Note that this approach ignores any differences in the recorded morphology.

The resulting additions (approximately 1000 entries) predominantly reflect cancer diagnoses in 2015 – 2017, but some are dated as far back as 1980, and the most recent is from 2019.

This was a one-off exercise, not expected to be repeated for Scottish data.

5.3. Break in coverage for Welsh participants

Cancer registry data for Welsh participants was originally provided by NHS England and its predecessors (see Appendix A for more information on evolution of data providers). However, NHS England ceased providing data for Welsh participants at the beginning of 2017. UK Biobank intends to incorporate enhanced cancer data from SAIL to resume coverage for Welsh participants in due course, using a similar method to that described in Section 5.2.

6. Data cleaning

Other than checking validity and linkage, there has been no detailed data cleaning, and no validation against other sources such as self-reported cancers has been attempted. One known issue is that there are a small number of participants with sex-specific cancer diagnoses not matching their sex.

These are likely to reflect typographical errors in the source data and we have generally made no attempt to remove these (or any other oddities from the data), except in a few cases:

- Records are removed where:
    - Cancer diagnoses are dated a long time after the participant's death
    - The records are clearly corrupted or incomplete (missing date or site)
    - The records contain demographic information (e.g. surname, date of birth) that contradicts UK Biobank participant records
- Morphology codes are redacted where either the histology or the behaviour component is missing or appears to be invalid (see Section 4.2 for more information). The rest of the record is left intact if no other problems are observed.

Appendix A

The following tables provide further information on data providers and record formats over time. Record formats correspond to the values in Data-Field 40019, indicating where a file format has changed (see Section 3 for more information).

Data providers for participants resident in England

Provider	Dates of receipt	Diagnosis dates	Record format
Medical Research Information Service, based at the National Health Service Information Centre (NHS-IC)	Feb 2011 – Jul 2015	Nov 1971 – Jun 2014	3
	Aug 2014 – Sep 2017	Jan 1971 – Dec 2016	23
National Cancer Intelligence Network (NCIN)	Aug 2014 only	Dec 1959 – Aug 2013	13
NHS England (NHSE)	Apr 2021 –	Mar 2015* –	73

* *Records before March 2015 were dropped from NHS England extracts to avoid conflicting duplicates between NHSE and NHS-IC records*

Data providers for participants resident in Wales

Provider	Dates of receipt	Diagnosis dates	Record format
Medical Research Information Service, based at the National Health Service Information Centre (NHS-IC)	Feb 2011 – Jul 2015	Nov 1971 – Jun 2014	3
	Aug 2014 – Sep 2017	Jan 1971 – Dec 2016	23
National Cancer Intelligence Network (NCIN)	Aug 2014 only	Dec 1959 – Aug 2013	13
NHS England (NHSE)	Apr 2021 only	Mar 2015* – December 2016**	73

* *Records before March 2015 were dropped from NHS England extracts to avoid conflicting duplicates between NHSE and NHS-IC records*

** *See Section 5.3 for more information*

Data providers for participants resident in Scotland

Provider	Dates of receipt	Diagnosis dates	Record format
Scottish Cancer Registry provided by the Information Services Division of NHS Scotland (ISD)	Oct 2012 – Nov 2016	Sept 1957 – Dec 2015	4
	Sep 2015 only	Mar 1985 – Aug 2014	22
Public Health Scotland (PHS) Enhanced cancer subset*	Aug 2018 and Oct 2020	Jan 1980 – Dec 2019	74
NHS Central Register, part of National Records of Scotland (NRS)	May 2018 –	Jan 1959 –	75

** See Section 5.2 for more information*