UK Biobank

Mortality data: linkage to death registries

Version 2.0

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

June 2020

This document details the data received from national death registries for UK Biobank participants.

Contents

1 Data providers

Data on UK Biobank participants who have died is currently received from NHS Digital for participants in England & Wales and from the NHS Central Register (NHSCR), part of the National Records of Scotland, for participants in Scotland.

All UK Biobank participants were flagged from the date of their recruitment into the study with these data providers.

New data are currently received from NHS Digital and NHSCR every month. Until recently this was every 6 months and 3 months respectively.

2 Data cleaning

Data is provided to UK Biobank with some personal identifiers which allow us to check that the record is for the correct participant. There has been a small amount of data with incorrect identifiers that have required correction.

Generally data is provided to researchers in the form that it has been received. One exception to this is to make minor adjustments to ICD-10 codes in cases where these have an extra superfluous digit. For example, some data has been received containing the “ICD-10 code” G200. (Note that all ICD-10 codes are received and made available without the decimal point.) However, there is no G20.0 in the UK ICD-10 system, as G20 has no subdivisions, and so this code has been corrected to G20.

Although we do not make substantive cross-referencing checks of the death data against other received datasets, some inconsistencies between datasets have been noted and the affected records set aside for further consideration. For example, there are a small number of cases where a hospital inpatient record ends after the participant's date of death; such records are investigated further before we decide whether they should be made available.

There are some cases where multiple (differing) death records are provided by the data providers, for example where a second death certificate is issued following a post-mortem. In such cases we make both records available, with the records each given a different "instance index" to distinguish them. (Note that this differs slightly from the usual use of instance index to label data items that are separated in time, since even where there are multiple death records they must agree on the date of death.)

3 Available data

3.1 Overview

The data UK Biobank receives from the death registry includes the date of death and the primary and contributory causes of death, coded using the ICD-10 system. Some records also have free-text cause of death information from the death certificate (see section 3.5 for further information).

Death data is available either to be downloaded as part of a main UK Biobank dataset or accessed via the Data Portal as described in sections 3.3 and 3.4 respectively.

3.2 Frequency of data updates

Data downloaded as part of a main UK Biobank dataset will only contain deaths that had been uploaded into our system as of the most recent Showcase Update, which typically occur two or three times per year.

Our intention is to update the death data available on the Data Portal much more frequently than that available through a Showcase Update. Researchers needing access to the most recent death data on a rolling basis should therefore use the Data Portal.

3.3 Data in a main dataset

The death data available in a main dataset consists of the Data-Fields in Category 100093 (apart from Field 40023). The main such fields are:

- Date of Death (Data-Field 40000)
- Underlying (primary) cause of death: ICD10 (Data-Field 40001)
- Contributory (secondary) cause of death: ICD10 (Data-Field 40002) – each cause is distinguished by having a different "array index" starting from 1 (this is a change to the format prior to June 2020 where they started from 0).
- Death record origin (Data-Field 40020) – either E/W for England & Wales or SCOT for Scotland (as per Data-Coding 1970).
- Death record format (Data-Field 40018) – a numerical code (see Data-Coding 261) indicating the format in which the data was sent to us. There are very few differences between formats which will be visible externally; one such difference is that free-text death information (see below) was only sent to us from Scotland in source 55, and not in the earlier Scottish sources (7, 19 & 54)

Category 100093 also contains a couple of additional Data-Fields that are not available through the Portal:

- Age at death (Data-Field 40007) – a derived data-field calculated as the interval between date of birth and the date of death.
- Description of cause of death (Data-Field 40010) – free-text cause of death information from the death certificate. This is only available for a subset of the data. See section 3.5 for further information.

3.4 Data on the Data Portal

Access to the death data on the Data Portal can be gained by adding Field 40023 to an application basket via the Access Management System (AMS). This gives access to the DEATH and DEATH_CAUSE tables as described below.

Note that any project which as of June 2020 has a basket containing a field from Category 100093 will be automatically granted access to the DEATH and DEATH_CAUSE tables. Hence, for such projects, it is not necessary to put in an Additional Data Request for Data-field 40023 in order to gain access to these Portal tables.

Please see the Accessing Data Guide on the Accessing your data page on Showcase for details of how to access the Data Portal.

3.4.1 Data Portal table structure

The death data on the Data Portal consists of two linked tables as shown in Figure 1.

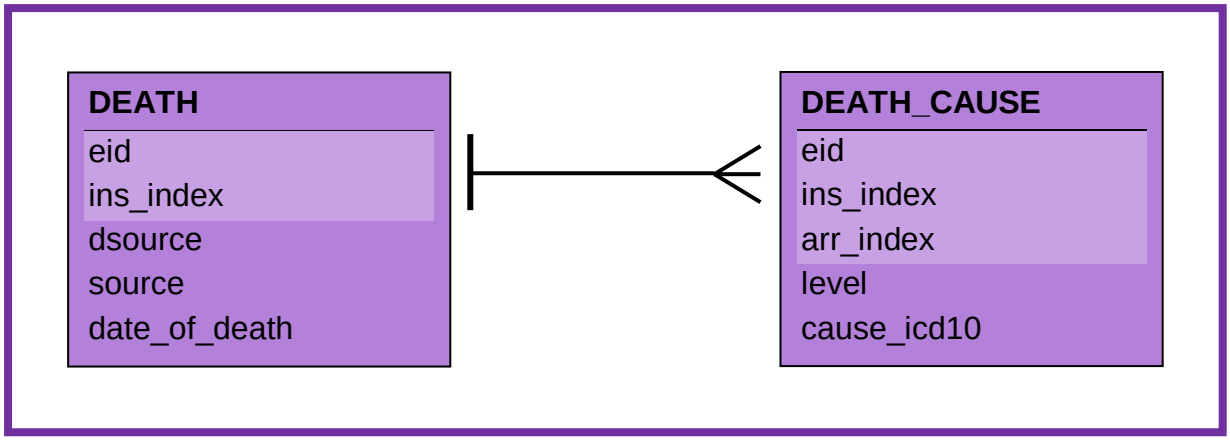

Figure 1: Structure of the Data Portal death tables

Each record in the DEATH table is uniquely identified by the eid (encoded identifier) of the participant and the instance index (ins_index) of the record. Participants usually only have one record in this table with ins_index = 0; however, as indicated in Section 2 there are a small number of participants for whom we have more than one death record.

The DEATH table includes the date that the participant died in the date_of_death field. This table also contains information on the data source (dsource) which is either E/W for England & Wales or SCOT for Scotland (as per Data-Coding 1970). The source field splits each of these further into a numerical code (Data-Coding 261) giving the particular data format in which this record was sent to us, but it is unlikely that most researchers will find this field useful.

The corresponding causes of death, coded using ICD-10, are provided on the DEATH_CAUSE table, with the eid and ins_index fields used to link back to the main record on the DEATH table. The arr_index is a sequential index, starting at 0, which labels each separate cause of death.

A primary cause of death is assigned level = 1 (and will have arr_index = 0) in this table and a contributory cause of death level = 2 (arr_index > 0).

All the Data Portal fields correspond directly to those available in a main dataset, with the same eid and ins_index, in the obvious way; e.g. date_of_death corresponds to Data-Field 40000 (Date of Death). Causes of death with arr_index = 0 (level = 1) correspond to those in Data-field 40001 (Underlying cause of death), and those with arr_index > 0 (level = 2) correspond to Data-Field 40002 (Contributory cause of death) with matching array index (so that array indices on Data-Field 40002 now start at 1).

Some examples of running SQL queries on the death tables are given in the next section.

3.4.2 Sample SQL

Tables on the Data Portal can either be downloaded in full, or queries can be run through the Portal to access specific subsets of data.

Some examples of extracting particular data-fields using SQL queries on the Portal are given below:

(1) Search for participant eids with primary cause of death D613 (i.e. D61.3):

```
select eid from death_cause
where cause_icd10 = 'D613' and level = 1
```

(2) As in (1) but allowing D613 to be a primary or contributory cause:

```
select eid from death_cause
where cause_icd10 = 'D613'
```

(3) As in (1) but extracting the date of death together with the participant eid:

```
select death.eid, date_of_death
from death_cause join death using(eid, ins_index)
where cause_icd10 = 'D613' and level = 1
```

(4) As in (2) but finding participant eids for which a cause of death is any ICD-10 code starting D61:

```
select eid from death_cause
where cause_icd10 like 'D61%'
```

Note that when using "like" to search for patterns, % represents zero, one or more characters.

3.5 Free-text cause of death

Some death records received by UK Biobank also contain the free-text cause of death that describe the sequence leading to death, i.e. the underlying cause and contributory causes leading to death.

However, this free-text information sometimes contains personal information about the participant. Therefore it is necessary for each record to be manually checked to remove any such information.

Due to the time required for this manual checking we have recently made the decision not to release free-text data for further participants for the foreseeable future. The data already processed is still available via Data-Field 40010 on Showcase (this Field is not available through the Data Portal).

The majority of death records from England & Wales have free-text cause of death up until the end of 2018, with the final such records in February 2019. Free-text cause of death was not provided from Scotland until recently, and so it is only available for records between January 2018 and August 2019.

Appendix A contains further details of the format of free-text cause of death data.

A Format of free-text cause of death data

The information from a death certificate is composed of the following two sections:

- Part I contains the disease or condition stated to be the underlying cause of death, and other significant conditions contributing to death but not related to the disease or condition causing it. It is organised into three (or four in Scottish death certificates) lines:
    - I(a) Disease or condition leading directly to death
    - I(b) Other disease or condition, if any, leading to I(a)
    - I(c) Other disease or condition, if any, leading to I(b)
    - I(d) Other disease or condition, if any, leading to I(c)

  Usually, the disease or condition that led directly to death is stated in line I(a) and any intermediate causes of death are stated in I(b), I(c) and I(d).

  In some cases the disease or condition that led directly to death is the same as the underlying cause of death. If this is the case, only line I(a) is completed. If death is due to an external cause such as a fall, this is provided as the underlying cause of death.

- Part II contains any other diseases, injuries, conditions, or events that contributed to the death, but were not part of the direct sequence leading to death.

An example of cause of death is the sequence:

- I(a) Post-transplant lymphoma
- I(b) Immunosuppression
- I(c) Renal transplant
- I(d) Glomerulonephrosis due to insulin dependent diabetes mellitus
- II Recurrent urinary tract infections

This would be processed into a single long string in Data-Field 40010 as:

1a. Post-transplant lymphoma; 1b. Immunosuppression; 1c. Renal transplant; 1d. Glomerulonephrosis due to insulin dependent diabetes mellitus; 2. Recurrent urinary tract infections

Death certificates sometimes contain "conclusion" or "verdict" sections at the end. Sometimes, due to the method by which this data is read from the death certificates, this final section cuts out mid-sentence, and so this information is only made available (at the end of the string above) when it appears complete.