UK Biobank

Cancer data: linkage from national cancer registries

Version 1.4

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

March 2013

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

Contents

1 Introduction

1.1: This manual details the procedure for linkage of UK Biobank participants to cancer registry data.

1.2: Data on cancer diagnoses for participants resident in England and Wales is provided to UK Biobank by the Medical Research Information Service, based at the National Health Service Information Centre (http://www.ic.nhs.uk/services/medical-research-information-service). The Information Services Division (http://www.isdscotland.org/Health-Topics/Cancer/), which is a part of NHS Scotland, provides UK Biobank with the cancer data records for participants resident in Scotland.

1.3: 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).

1.4: 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 documentation on prevalent and incident cases in the Data Showcase for more details about the numbers of cancers currently in the database.

1.5: While we have attempted to provide meanings for all codes supplied here, there has been no detailed data cleaning. For example, there are a small number of participants with apparent duplicate data (i.e. diagnosed with the same cancer type on the same date, or shortly thereafter) and we have made no attempt to remove these (or any other oddities from the data). The Cancer Outcomes Working Group is preparing a more refined dataset for cancer outcomes, which will be available in due course.

2 Data collected

2.1: Data from the Information Centre and Information Services Division is sent to UK Biobank quarterly.

2.2: The data presented in Showcase and available to researches comprises:

- Date of cancer diagnosis
- Age at cancer diagnosis
- Type of cancer: ICD-10
- Type of cancer: ICD-9
- Reported occurrences of cancer
- Histology code
- Behaviour code

2.3: The type of cancer is coded according to the International Classification of Diseases, 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 amount of cancers was diagnosed before 1979, when the eighth revision ICD (ICD-8) codes were in use. (However, all of these particular ICD-8 codes were retained when using ICD-9 format; hence, we have grouped these cancers into 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 current version is ICD-10 4th edition, which was implemented in 2010. The ICD-codes are presented in a tree-like structure, grouped according to ICD-10 chapter order (rather than by the diagnosis name). Further details on this coding system are available at http://apps.who.int/classifications/icd10/browse/2010/en.

2.4: Please note that there will be a time-lag between the date of cancer diagnosis and the date that the Information Centre (& thence UK Biobank) receive the cancer registration and incorporate this data in the Resource (usually 6-12 months delay, but sometimes longer).

2.5: The histology and behaviour codes of neoplasms are represented as separate variables. These variables are presented as five-digit codes in ICD10-O-3 (1), ranging from M-8000/0 to M-9989/3. The first four digits code the histology and the fifth digit codes the behaviour. The coding translations are found in a separate document in the 'Additional Resources' tab.

2.6: The morphology (behaviour) code relates to the 5th digit (i.e. after the slash) and is coded as follows:

/0 Begin

/1 Uncertain whether benign or malignant

- Borderline malignancy
- Low malignant potential

/2 Carcinoma in situ

- Intraepithelial
- Noninfiltrating
- Noninvasive

/3 Malignant, primary site

/6 Malignant, metastatic site

- Malignant, secondary site

/9 Malignant, uncertain whether primary or metastatic site

2.7: For participants with more than one type of cancer, data relevant to each cancer type is available. Hence, all of the cancer data-fields are stored as 'variable instance' in the Data Showcase, with each occurrence of cancer presented as a separate 'instance'.

References

1. Fritz A, Percy C, Jack A, Shanmugaratnam K, Sobin L, Parkin DM, et al. International Classification of Diseases for Oncology, Third edition (ICD-O-3): World Health Organization; 2000.