UK Biobank

Validation and cleaning of externally collected data

Version 1.0

http://www.ukbiobank.ac.uk/
July 2014

biobank uk
Improving the health of future generations

Contents

1. Introduction

1.1 UK Biobank has linked to a number of externally collected data sources (e.g. death and cancer registries, and hospital episode statistics) in order to obtain comprehensive follow-up information for its half a million study participants.

1.2 Integrating data from external sources into UK Biobank's database (and ultimately presenting this information in Data Showcase) is a multistep process. This process involves obtaining, documenting, handling, and storing large and complex data items; further detail of which can be found here.

1.3 Ambiguities in the data can arise during processing, such as invalid clinical classification codes, implausible date values or mismatches of participants' records. Defining cleaning rules for handling these is essential in order to provide high-quality data for research purposes.

1.4 This aim of this document is to outline UK Biobank's approach to validating and cleaning data obtained from record linkage sources.

2. Data receipt, inspection and documentation

2.1 On receipt of the data file from the external data provider, the contents are inspected to understand exactly what information is contained. In particular, the format and values of individual data fields are scrutinised to understand whether they conform to those indicated in the data dictionary, and whether the information is useful for research purposes. Any coding ambiguities identified at this stage are clarified by UK Biobank's data analysts or with the data provider, if necessary.

2.2 A document is prepared specifying the criteria for importing the data into UK Biobank's database. This includes where the data field is located in the source file; the data field name and description; the regular expression of the data field (i.e. its format) and whether there is a definitive list of values (i.e. it follows a coding system). It is also noted whether the data field will be imported into the database for internal and/or external purposes and if so, its storage location.

3. Data import and validation

3.1 For each data file a bespoke program (developed by the Clinical Trial Service Unit and Epidemiological Studies Unit (CTSU), University of Oxford) is used to import the data into UK Biobank's database, based on criteria specified by the data analysts in the import documentation (as described in section 2.2).

3.2 All data imported into the UK Biobank database are validated using the following checks:

i. Checks for mismatches

The accuracy of the matching algorithms used by the external data providers (details of which are provided here) is checked on a record-by-record basis (i.e. to assess the likelihood that the data belongs to the UK Biobank participant it has been linked to). This check is performed by comparing items of identifiable data (such as name, date of birth, NHS number), which UK Biobank collected for each participant at recruitment and which is also contained in the data file supplied to UK Biobank.

The mismatch rate is estimated to be <0.1%, largely due to a very high proportion of the cohort having a NHS number (or CHI number in Scotland), which acts as a unique identifier for linkage purposes.

ii. Checks for data formatting

A regular expression filter is used to check that the values for a data field are in the format specified in the import documentation. For instance, a data field which is expected to contain a date value would be checked to see if it follows the expression dd-mm-yyyy, while a data field which is expected to contain 3 letters followed by 3 numbers would be checked to see if it follows the expression aaannn.

iii. Checks against a definitive list of coded values

Checks are performed on data fields where the value must comply with those specified in a definitive list according to a data dictionary. For instance, a data field which is expected to contain a date value would be checked to see if the date is actually within a plausible range, and a data field which is expected to contain a clinical classification code (such as ICD10), would be checked to see if the code is valid.

4. Data cleaning

4.1 Values which fail validation are flagged for attention and investigated further until a decision is made about whether to exclude the record from import (i.e. the record does not belong to a UK Biobank participant), to modify the list of definitive values (i.e. the data dictionary) to incorporate a new valid code, or to modify the value into a valid code.

4.2 To view the data cleaning rules that which have been applied to data files obtained from external sources, please see the data cleaning section of the essential information page.