UK Biobank

Integrating electronic health records into the UK Biobank Resource

Version 1.0

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

January 2014

This document provides details on the procedures involved in processing and integrating electronically linked health-related records into the UK Biobank central data repository.

Contents

1 Aim 2
2. Types of linked electronic health records in UK Biobank 2
3. General approach 3
4. Linkage data tasks 4

1. Aim

This document provides details on the procedures involved in processing and integrating electronically linked health-related records into the UK Biobank central data repository.

2. Types of linked electronic health records in UK Biobank

Currently, there are three different types of health-related outcome records that have been (or are in the process of being) incorporated into the central database.

Type of data	External provider	Region	Period of data available
Deaths	HSCIC	E&W	April 2006 onwards
	ISD	Scotland	
Cancer registrations	HSCIC	E&W	since inception - 1980s
	ISD	Scotland	since inception – 1950s
Hospital inpatient episodes	HES (HSCIC)	England	since inception - 1997
	PEDW (SAIL)	Wales	since inception - 1999
	SMR	Scotland	since inception - 1981

HES: Hospital Episode Statistics; HSCIC: Heath & Social Care Information Centre; ISD: Information Services Department; PEDW: Patient Episode Data for Wales; SAIL: Secure Anonymised Information Linkage; SMR: Scottish Morbidity Records

Other potential linkages:

Plans are underway to include primary care records, dental records, pharmacy dispensing records, imaging records, screening programmes, disease-specific registries (e.g. for myocardial infarction, surgical interventions), social services, education and environment records (where available), in the medium to long term.

3. General approach

The process of incorporating routine electronic health records into the UK Biobank resource is complex due to the variety of external providers, and hence variety of data content and format (which may change over time). The approach to integrating data from external source data providers (e.g. deaths, cancers, hospital episodes and primary care records) is summarised in Figure 1, and a brief overview of each step is described below.

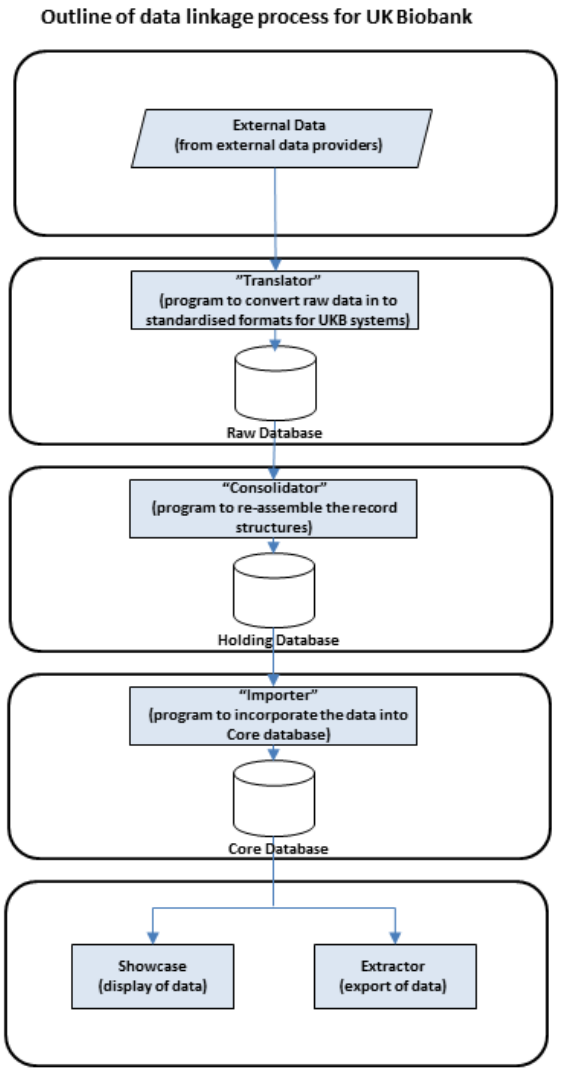

Figure 1. Schematic diagram of how electronically linked data from external data providers is processed into UK Biobank.

For each file, a "Translator" program (developed by the Clinical Trial Service Unit and Epidemiological Studies Unit (CTSU), University of Oxford) converts the data into a standard format that is suitable for integration in the internal UKB databases held at CTSU. These data are then reorganised or consolidated (with the use of in-house "Consolidator" program) and imported into the main ("Core" database). The data can then be displayed externally via Showcase and subsequently exported for approved research projects.

4. Linkage data tasks

There are several steps involved in processing the raw data file in order to make it publically available. Figure 2 shows the generic IT tasks involved in the processing of each linkage data file.

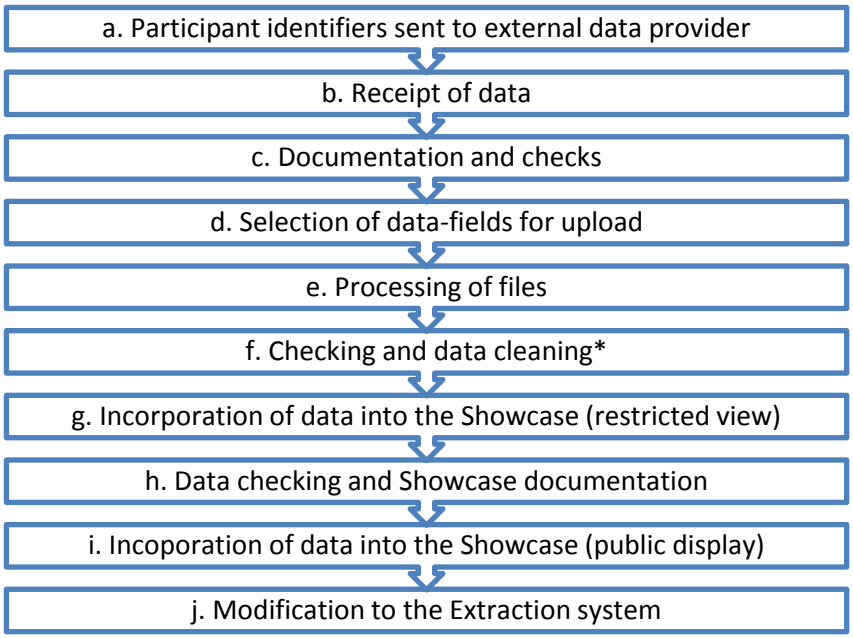

Figure 2: The steps involved for the incorporation of externally linked data into the UK Biobank central database
***Information on UK Biobank's data cleaning procedures are detailed elsewhere and can be found on the website.**

a. Participant's identifiers sent to external data provider

All of the processes (e.g. applications, permissions) required to obtain the data from external organisations are managed by UK Biobank. At the point at which agreement has been reached, UK Biobank (CTSU) provides appropriate participant identifiers (unique participant identifier (PID), NHS number, date of birth, gender and postcode) for matching purposes. The data record for each matched individual is then (usually) assigned an accompanying pseudo-identifier and the data file returned to UK Biobank (CTSU). More information about the matching algorithms used by external data providers is provided elsewhere and is available on the UK Biobank website.

b. Receipt of data

The data are received via a secure connection and stored on an encrypted server with restricted access. During this step the identifiers sent to the external provider are checked against the data received. If there is a mismatch for any of the fields, the record is flagged and investigated further. Files are sent to UK Biobank on an agreed basis (e.g. quarterly for death and cancer records; annually for hospital activity data).

c. Documentation and checks

On receipt of the data in CTSU, the contents of the file are evaluated (size, format, field names, values) in accordance with the data dictionary and documented. Any potentially problematic/ambiguous data-fields are identified for further investigation (or exclusion). Decisions and suggestions are also made where there is a need for a derived variable to be generated.

d. Selection of data fields

A list of data-fields to be processed is generated, although not all of these data-fields are finally incorporated into the central (Core) database, as many may be redundant or considered uninformative for research purposes.

e. Processing of files

The program used to process the files is modified in accord with the accompanying data documentation in order to upload and process the file into a designated database. Data ambiguities are clarified with the data analysts and external provider(s).

f. Checking and data cleaning

A random sample (without repetitions) is drawn from the data (usually 100 records) and tested for data integrity and that it matches the source information. Particular attention is given to date and free-text fields to ensure the formatting and truncation are correct.

Minimal data cleaning is performed (please refer to the “Data Cleaning” document for further details) and consists largely of ensuring the coding is within recognised parameters.

For the death registry files, data cleaning of free-text entries denoting cause(s) of death is required to remove potential identifying information.

g. Incorporation of data into the Showcase (restricted view)

The approved data fields are consolidated and incorporated into the central internal database, accessible only to internal UK Biobank staff. Additional programming may need to be done at this stage to merge or create derived variables. For example, data on hospital activity from the 3 separate data providers (England, Wales and Scotland) is amalgamated, where possible.

h. Data checking and Showcase documentation

The data-fields are checked to ensure the import has proceeded correctly and explanatory documentation for each data-field is uploaded to aid researchers in the derivation and interpretation of each data item.

i. Incorporation of the data into Showcase (public display)

Once all checks are completed, the data is made publically available via the Data Showcase.

j. Modification to the Extraction system

In some circumstances, the extractor software may need to be modified to allow the appropriate data to be exported to external researchers (for example, to allow modifications to the internal data dictionaries).