UK Biobank

External Data Providers: Matching Algorithms

version 1.0

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

biobank uk
Improving the health of future generations

1 Introduction

1.1 Integrating routine health records and registry data into the UK Biobank resource provides a unique opportunity to better understand the causes of a wide range of health conditions.

1.2 Death registrations, cancer registrations and hospital episode data are currently being obtained by UK Biobank on a regular basis and are being made available to the wider research community, via Data Showcase.

1.3 National coverage (89% of participants recruited in England; 7% in Scotland; 4% in Wales) requires linkage to multiple data providers, with a range of different access procedures and requirements for matching participants.

1.4 This document provides details of the linkage process for each data provider currently providing linked data to UK Biobank, with particular focus on the matching algorithms used to link participants at an individual level between UK Biobank and externally provided datasets.

2 External Data providers

2.1 UK Biobank works closely with multiple data providers, including the Health and Social Care Information Centre (HSCIC) in England, the Information and Statistics Division of National Services Scotland, and the Secure Anonymised Information Linkage (SAIL) databank in Wales.

2.2 Datasets have different inception and censoring dates (e.g. hospital admissions data in Scotland, which is available from 1981, compared to hospital admissions data in England, available from 1996) and this presents potential problems in terms of determining completeness of retrospective coverage and of prospective follow-up.

2.3 Data providers have different criteria for matching participants, and require UK Biobank to provide different items of identifiable data. Note that all data exchanged between UK Biobank and external data providers is done with high levels of data security. UK Biobank operates under the UK Biobank (CTSU) System Level Security Policy, which is implemented by the IT Security Manager.

2.4 In general, participant identifiers for the **entire cohort** are transferred securely to data providers operating safe havens for data linkage in order to capture cross-border activity (e.g. a participant who is resident in England, but is treated in a hospital in Scotland). Regular communication between data providers ensures that any cross border activity is automatically notified to the country where the participant is resident.

2.5 Once the linkage is in place, UK Biobank receives a onetime ‘period linkage’ of all available data held by the data provider. Updates to the data are then received at variable frequency, as agreed between UK Biobank and each data provider, usually annually or quarterly.

2.6 Where possible, UK Biobank obtains a match rank score for each participant identified, so that the quality of the matching procedure can be assessed.

2.7 Table 1 provides details of the linkage process for each separate data provider providing linked data to UK Biobank, including the type, frequency and years of availability of the data received, the identifiers provided by UK Biobank, and the matching algorithm involved in the linkage. Note that for some data providers, identifiers were sent out in smaller batches over a period of time (depending upon the capacity of each data provider’s systems). Also note that in each linkage the total number of participants may differ, due to participant withdrawal from the study over time.

Table 1: External data providers

Data Type	Data Provider	Coverage	Frequency of updates	UKB identifiers sent to data provider	Matching algorithm for Linkage
Death and Cancer Registrations	Health and Social Care Information Centre (HSCIC)	England and Wales (death data prospective from 2006, cancer data available from ~1970)	Quarterly	➢ NHS number Participant ID surname forename date of birth sex address postcode ➢ Identifiers sent for participants resident in England and Wales	1. **NHS number**, **date of birth** *single match – accept, no match – go to step 2* 2. **NHS number**, 2/3 parts of **date of birth**, first 3 characters of **surname**, first character of **forename** *single match – accept, no match – go to step 3* 3. 2/3 parts of **date of birth**, **surname**, first 2 characters of **forename**, **postcode** *single match – accept, no match – go to step 4* 4. 2/3 parts of **date of birth**, **surname**, **forename** *single match – go to step 5, multiple matches/no match – no trace* 5. **date of birth**, **surname**, **forename** *single match – accept, no match – no trace*
	Information Services Division (ISD), National Services Scotland	Scotland (prospective from 2006)	Annually	➢ participant ID forename surname date of birth sex NHS number postcode ➢ Identifiers sent for all participants	<u>Exact match using:</u> 1. **CHI number*** *single match – accept, no match – go to step 2* <u>Fixed threshold probability matching using:</u> 2. **sex**, **date of birth**, **surname** soundex**, **forename**, **CHI number***, **postcode**

					Each identifier is scored on how well the information compares, i.e. an exact match is awarded a full score, while partial matches are awarded a lower score. A threshold of 32 points was set for UK Biobank participant matching. * NHS numbers provided by UK Biobank were matched directly with the CHI database. ** Surnames and maiden names were phonetically transformed using the New York State Identification and Intelligence System (NYSIIS) and then compressed using a soundex algorithm.
Hospital Admission data	Health and Social Care Information Centre (HSCIC)	England and Wales (from 1996)	Annually	➢ participant ID NHS number sex date of birth postcode ➢ Identifiers sent for participants resident in England and Wales	<u>Years 1996-97 to 2009-10:</u> 1. **sex**, 2/3 parts of **date of birth**, **NHS number** *single match – accept, no match – go to step 2* 2. **sex**, 2/3 parts of **date of birth**, **postcode**, **local patient identifier** *single match – accept, no match – go to step 3* 3. **sex**, full **date of birth**, outcode of **postcode** (full postcode required if date of birth equals 01 January) *single match – accept, no match – no trace* <u>Years 2010-11 to present:</u> 1. **sex, date of birth, NHS number, postcode** *single match – accept, no match – go to step 2* 2. **sex, date of birth, NHS number** *single match – accept, no match – go to step 3* 3. **sex**, 2/3 parts of **date of birth**, **NHS number**, **postcode** *single match – accept, no match – go to step 4* 4. **sex**, 2/3 parts of **date of birth**, **NHS number**

					single match – accept, no match – go to step 5 5. **NHS number, postcode** *single match – accept, no match – go to step 6* 6. **sex, date of birth** (excluding 01 January), **postcode** (excluding communal establishments e.g. hospitals, prisons and army barracks), where **NHS number** does not contradict the match (i.e. NHS number is present in one dataset but is null in the other, or is null in both datasets) *single match – accept, no match – go to step 7* 7. **sex, date of birth** (excluding 01 January), **postcode**, where **NHS number** does not contradict the match (i.e. NHS number is present in one dataset but is null in the other, or is null in both datasets) *single match – accept, no match – go to step 8* 8. **sex, date of birth** (excluding 01 January), **postcode** *single match – accept, no match – no trace*
	Information Services Division (ISD), National Services Scotland	Scotland (from 1981)	Annually	➢ participant ID forename surname date of birth sex NHS number postcode ➢ Identifiers sent for all participants	Exact match using: 3. **CHI number*** *single match – accept, no match – go to step 2* Fixed threshold probability matching using : 4. **sex**, **date of birth**, **surname** soundex**, **forename**, **CHI number***, **postcode** Each identifier is scored on how well the information compares, i.e. an exact match is awarded a full score, while partial matches are awarded a lower score. A threshold of 32 points was set for UK Biobank participant matching. * NHS numbers provided by UK Biobank were matched directly with the CHI database.

					** Surnames and maiden names were phonetically transformed using the New York State Identification and Intelligence System (NYSIIS) and then compressed using a soundex algorithm.
	Secure Anonymised Information Linkage (SAIL) Databank	Wales (from 1998)	Annually	➢ participant ID NHS number ➢ Identifiers sent for all participants	1. **NHS number*** * Only exact matches were returned to UK Biobank.