UK Biobank

Home location history

Version 1.1

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

May 2025

This document details the procedure for the derivation of participant home location history

Contents

1. Understanding the dataset

From the time that each participant was invited to join the study, up to the present day, UK Biobank has maintained a record of their current residential address in its contact database.

The addresses themselves are not disclosed to researchers owing to their identifying nature; only designated members of staff have access to this information. The addresses have instead been geocoded and the resulting coordinate pairs, whilst still restricted, will be released to researchers where necessary for their project.

The published data consists of Easting/Northing coordinate pairs, using the British National Grid projection, each with an associated date. There are three resolutions available: 100m, 1km and 10km. Researchers who request these data are encouraged to use those at a lower resolution (1km or 10km), which will be of sufficient granularity for most research purposes. Grid co-ordinates to a 100m resolution will only be released for specific research projects that fully require this level of granularity, and via a special release procedure. Researchers should contact the Access Management Team to discuss this before deciding whether to proceed.

1.1. Address data collection

The initial address for each participant is the one to which their invitation to join UK Biobank was posted. The current address is verified with the participant at each assessment visit.

Participants are encouraged in newsletters and other communications to check their address and other contact details, and can either update these online or contact the Participant Contact Centre when they have moved to a new location. In addition, a number of exercises have been undertaken by UK Biobank to bring the contact database up to date. These exercises make use of data sourced from the NHS or other third parties to update addresses for participants whose address is known or suspected to be outdated.

The date that UK Biobank was notified of each address is recorded alongside its location. Please note that these dates are expected in general to lag significantly after the date of the actual move.

1.2. Address data cleaning

All addresses received by UK Biobank are validated and standardised using the software Experian QAS Batch. Where an address cannot be matched by this software, it is manually reviewed and the participant is contacted to request a correction if necessary.

When a new address is entered, UK Biobank has no mechanism for recording whether it represents a change of address, or a correction to the address. Minor adjustments to an address are often received, for example adding the house name. The process used for determining whether two addresses are the same is to compare the standardised version of each address. In some cases, addresses have been standardised multiple times with slightly different results due to changes in the reference data used by QAS Batch and/or changes in the software configuration. If two addresses have any matching results amongst these standardised versions, then they are deemed to be the same.

Where two consecutive addresses match, the second instance of the address is omitted from the address history.

Where an address was valid for less than 7 days, this is assumed to be an error and the address is omitted from the address history. Inspection of these has revealed that most are data entry errors that are fixed later.

In some circumstances, the last known address may be marked invalid, without a new address being added. This happens when UK Biobank receives an item of mail marked 'return to sender' or occasionally when a participant removes their details from the participant portal. In such a case, an entry appears in the address history giving the date of the event and a pair of missing coordinates designated by the value -998 (Data coding 2535).

1.3. Derivation of geo-coordinates

The software used to geocode the UK Biobank addresses, originally at Leicester University and later internally within UK Biobank, is Experian QAS Batch. The QAS software geocodes addresses by verifying them against the official (Royal Mail) postal address files for the UK. Cleaned records are then assigned a match result based on the accuracy of the original address. The same software package, using the Experian DataPlus dataset "United Kingdom Location Essential" (GBRGEO), then attempts to determine precise (1m) geocoordinates for each verified address. Only results for addresses marked as Verified Correct or Good Full Match (match code "R9") were accepted.

The 100m coordinates were then derived from the 1m coordinates output by QAS Batch, by dividing by 100, truncating any decimal part, and multiplying by 100, and a similar process was used for the 1km and 10km coordinates. For example, a 1m-resolution coordinate of 567890 would result in a 100m-resolution coordinate of 567800, a 1km-resolution coordinate of 567000, and a 10km-resolution coordinate of 560000. This means that the given coordinates correspond to the southwest corner of a grid square of the relevant size, containing the address.

Overseas addresses, and a small proportion of those within Great Britain, cannot be allocated grid coordinates and these are given a special value of -999 to indicate that UK Biobank has been informed that a change of address took place, but cannot provide coordinates.

There are also cases where the grid co-ordinates corresponding to home locations fall within low population density areas, i.e. where fewer than 50 individuals live within a 1km^2 circular area centred on the home location. In these cases, the grid coordinates are redacted using a special value of -997. It is recommended that researchers use both sets of fields to get a complete dataset.

Special values for these datasets are described in Data-Coding 2535.

1.4. Grid coordinate system

The grid coordinate data are provided in the British National Grid (i.e. OSGB 1936) projection. The EPSG code for this projection is 27700.

The grid coordinate measurements refer to easting and northing with a reference point near the Isles of Sicily. The coordinates specify the south-west corner of the grid square, of the relevant resolution, which contains the participant's address.

The values are given in the units of meters, and there are no negative values (as there are in the case of latitude/longitude). Negative values in the UK Biobank data indicate special values as given in Data Coding 2535.

2. Available data

Field id	Field name	Notes
32220	Home location history - date first recorded	Date that UK Biobank became aware of this address
32221	Home location history - east co-ordinate (100m resolution)	Highly restricted. To be used in conjunction with north co-ordinate (field 32222)
32222	Home location history - north co-ordinate (100m resolution)	Highly restricted. To be used in conjunction with east co-ordinate (field 32221)
32223	Home location history - east co-ordinate (1km resolution)	Restricted. To be used in conjunction with north co-ordinate (field 32224)
32224	Home location history - north co-ordinate (1km resolution)	Restricted. To be used in conjunction with east co-ordinate (field 32223)
30066	Home location history - east co-ordinate (10km resolution)	Restricted. To be used in conjunction with north co-ordinate (field 30067)
30067	Home location history - north co-ordinate (10km resolution)	Restricted. To be used in conjunction with east co-ordinate (field 30066)

The data items are linked together using array indices. For example, the date value in field 32220 with array index = 1 corresponds to the east and north coordinates which also have array index = 1 for the same participant. Instance indexing is not used for this data.

Please note that requesting access to grid coordinate data in this category will preclude access to the census/administrative area fields in Category 703, due to the risk of reidentification. Researchers should carefully assess whether they require grid coordinate data or census/administrative area data for home locations before requesting access.

3. Data quality

As noted above, the published data corresponds to the contact addresses held by UK Biobank for its participants. There are limitations with this system including:

- There is no mechanism for backdating an address, e.g. if a participant informs us today that they moved last year, the recorded date will be today
- There is no mechanism for distinguishing between a new address and a correction, e.g. if a house number was incorrectly entered as 21 instead of 12, and later corrected, both addresses, with potentially different coordinates, will persist in the history

- The method for determining whether two addresses are the same (described above) results in false negatives in some cases. In such cases the grid coordinates would be correct, but a single participant address would be split between two separate entries in the history
- There are many participants who are no longer in contact with UK Biobank, or who have not kept their contact data up to date. Some of these will no longer live at the last recorded address, and/or may have an incomplete location history.