Summary de-identification protocol V2

1. Introduction

1.1 This note sets out the UK Biobank policy for the de-identification of participant data (*Participants and Participant Data)* prior to its release to researchers. At the outset, two clear distinctions should be made:

1.1.1 This note ***does not*** specifically address the simple fact of an individual being identified as a Participant in UK Biobank, as many Participants choose (quite reasonably) to volunteer this information about themselves. However, UK Biobank does not itself release the identity or confirm the identity of any Participant.

1.1.2 This note ***does*** specifically address the release of Participant Data by UK Biobank to researchers where the manner in which the Participant Data is released could inadvertently identify a Participant (bearing in mind other information about the Participant which may already be in the public domain).

1.2 To illustrate: the fact that *Mr Jones* is a participant in UK Biobank is significant but not, of itself, an overtly significant item of information. Nevertheless, releasing information about the state of Mr Jones health to researchers in such a way that a) Mr Jones is (or can be) identified and then b) cross-referenced directly to the UK Biobank phenotypical or genotypical information which UK Biobank holds on him, would be highly problematic. The purpose of this note is to set out the steps that UK Biobank takes to, as far as UK Biobank possibly can, ensure that this does not happen.

1.3 This note also sets out certain details about the nature of the Participant Data which UK Biobank holds on Participants, UK Biobank's legal obligations, considerations around health-record linkage, practical considerations about identification and finally the actual de-identification protocols.

2. Background

Identifiable Participant Data

2.1 The Participant Data that UK Biobank holds includes the following identifiable data on each Participant (*Identifiable Participant Data)*: name, address, email, mobile phone number, NHS number. This Identifiable Participant Data is ***not*** released to researchers[1].

2.2 This Identifiable Participant Data (the minimum amount required) is only released to third parties where strictly necessary and always subject to a compliant legal agreement: for example, where UK Biobank needs to send a COVID-19 testing kit or its newsletter to participants it has to provide a suitable mailing list (with names and addresses) to the distribution centre.

2.3 Within UK Biobank, this identifiable data is stored and encrypted and only accessible by a very small number of UK Biobank personnel who ***need*** to access it.

De-identified Participant Data

2.4 All the other Participant Data that UK Biobank holds, which includes the detailed research data on each Participant is stored in a de-identified manner (*De-identified Participant Data*), is categorised below:

2.4.1 measures of the Participant's phenotype, such as height, weight and blood pressure (approximately 2,000 phenotypes per Participant, as further detailed here http://biobank.ndph.ox.ac.uk/showcase/schema.cgi?id=1);

[1] Please also see section 6.5 below.

2.4.2 measures of the Participant's genome, this includes genotype, exome sequence and whole sequence data;

2.4.3 biomarkers created by assay of the Participant's samples, which include common biomarkers (such as cholesterol), infectious disease markers, proteomic and metabolomic markers;

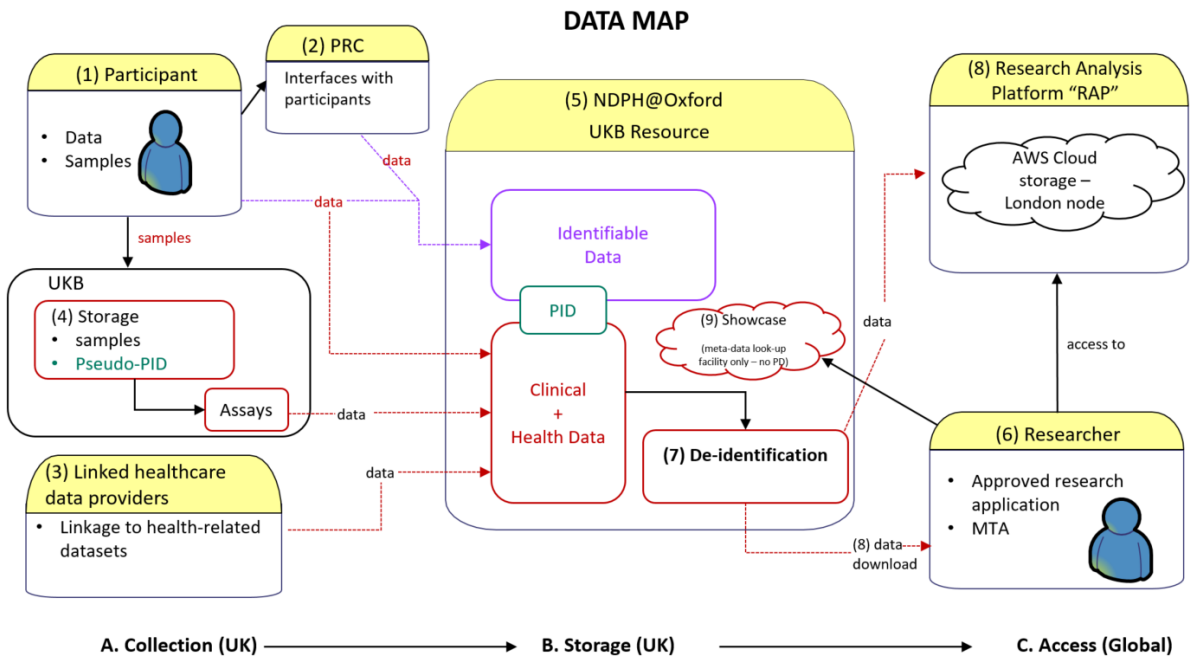
2.4.4 imaging data (on up to 100,000 Participants) as the result of MRI scans of the head, the heart and the body, plus ultrasound and DEXA; and

2.4.5 data derived from health record linkages including hospital records, primary care records, death and cancer registries or any other sources of clinical data; and

2.4.6 other special category data derived from baseline/online questionnaire responses and interviews, such as past illness / disease history, dietary, cognitive and physical measures

Data Storage and PIDs

2.5 The UK Biobank IT systems store the Identifiable Participant Data – see diagram below – separately from the main databases where the De-identified Participant Data is stored. Further, each participant is assigned a unique identifier known as a PID, which is randomly generated from the participant's NHS number and cannot be reverse engineered by third parties. The PID is stored with both the Identifiable Participant Data and with the De-identified Participant Data and acts as the "primary-key" to link the information together.



3. UK Biobank's legal obligations

3.1 As part of the consent process, UK Biobank explicitly undertook to (a) preserve the anonymity of its Participants and more critically (b) not to release data about Participants in such a way that Participants can be identified. These undertakings apply to both living and deceased Participants.

3.2 UK Biobank legal obligations mirror these undertakings provided to Participants during the consent process. These legal obligations are as follows:

3.2.1 UK Biobank owes an obligation of confidentiality to the Participants not to identify them. There are no de facto privacy issues (under the Human Rights Act) as "privacy" as a legal concept only applies to identifiable individuals; and

3.2.2 UK Biobank owes certain duties to participants under the GDPR, which are set out in more detail here https://www.ukbiobank.ac.uk/explore-your-participation/basis-of-your-participation [2].

3.3 UK Biobank ensures that researchers comply with these obligations relating to Participant Data – bearing in mind that only De-identified Participant Data is ever released - through the mechanism of its Material Transfer Agreement (*MTA).* The MTA has recently been revised and updated to take into account changes in practice and regulation https://www.ukbiobank.ac.uk/media/5cclro0y/applicant-mta-data-only-2021.pdf .

3.4 The De-identified Participant Data is only released to researchers (whose application has been approved by UK Biobank) who agree to enter into UK Biobank's MTA, which sets out in detail what researchers are entitled to do with the De-identified Participant Data. The MTA also contains specific provisions prohibiting researchers from trying to take any steps to pro-actively identify or contact Participants.

4. Health record linkage

4.1 As a prospective resource, UK Biobank follows up Participants through the health records (primary care, secondary care and others): it has explicit consent to do this https://www.ukbiobank.ac.uk/media/05ldg1ez/consent-form-uk-biobank.pdf .

4.2 These health record linkage arrangements are always covered by a specific agreement with the relevant counterparty (such as NHS Digital) and UK Biobank has a policy of seeking to conduct, wherever possible, the linkage such that Participants are not re-identified during the course of the linkage.

5. Practical considerations

5.1 To recap, as UK Biobank has an obligation to provide De-identified Participant Data to researchers in a manner which (a) preserves the anonymity of its Participants and (b) as far as practically possible does not enable Participants to be inadvertently identified.

5.2 There are certain items of information within the Participant Data, which have varying degrees of potential to identify a Participant (either alone or in combination). These direct identifiers are removed from the Participant Data and these include date of birth; gender; name of GP; ethnicity; postcode; certain specific event instances (such as admission of a participant to a particular hospital); potentially identifying codes (such as rare occupation) and unedited free text fields in linkage health-record data relating to the Participant.

5.3 These items vary in their potential to identify Participants, with unedited free text fields and detailed post code being the most powerful. In combination, certain of these items (for example post code, gender and date of birth) can serve to significantly increase the risk of re-identification.

5.4 In terms of the NHS number, which is stored by UK Biobank as an encrypted PID, each dataset released to a researcher contains project-specific encrypted identifier (EID) to replace the PID. An EID is a number algorithmically generated by UKB specifically for that particular research application. The link between a) the NHS number and the PID and in turn b) the PID and the EID is securely stored by UK Biobank.

5.5 There are also certain data items which are inherently unique to a Participant – for example genetic sequence data. However, the re-identification risk posed by this type of data is in practice relatively small. Using the sequence data as an example, a researcher in possession of sequence data (or a collection of SNPs or tandem repeats) would have to possess (a) another comparable genetic sequence of the

[2] UK Biobank is not subject to the Freedom of Information Act 2000 and comparable upcoming legislation (as it is not a public body).

Participant from a source which identified the Participant as well as holding the genetic sequence of the Participant from UK Biobank and (b) the computing systems to match the two sequences. This is technically possible, but the actual risk of re-identification is *in practice* relatively small.

5.6 UK Biobank is conscious that technology in this area moves at a considerable speed and is particularly conscious that there are certain open access resources (particularly those used for generating ancestral trees from genetic data) where individuals can upload their own genotype or sequences. Actions such as this do serve to increase the possibility of re-identification and in this regard further guidance to Participants is under consideration.

6. UK Biobank's protocol

6.1 Participant Data will always be released to researchers with distinct encrypted random numbers. In other words, each set of Participant Data relating to an individual Participant is identified by an encrypted random number which has been generated specifically for each application: this is the EID which is generated from the PID).

6.3 Further, UK Biobank does not release:

6.3.1 potentially directly identifiable information such as full DOB and instead releases just month and year of birth, any scanned images which contain identifiable facial images and any unedited free text fields from health-records (as these could contain identifiable references to the participant or other individuals); and/or

6.3.4 potentially indirectly identifiable information such as detailed post codes, where UK Biobank's default position is to only provide details of an address (based on grid references or postcodes) to 1km accuracy (equivalent to the last 3 digits of a postcode being removed).

6.4 Where researchers wish to use data fields which contain more potentially identifiable information for an approved research purpose, this is considered by UK Biobank on a case-by-case basis, and appropriate methods are adopted to protect participant anonymity.

6.5 By way of example, if a researcher requested detailed address data for the purposes of linkage to an environmental dataset (e.g., air pollution estimates), UK Biobank will release the data in a phased manner such that is not possible to link the address data with any other data-field. Hence, the address data is released first with one set of EIDs, and only once the linked data is returned to UKB do the researchers receive the full dataset (excluding address data) with a different set of EIDs.

6.6 UK Biobank would make two final caveats about this de-identification protocol:

6.6.1 it is not a fail-safe guarantee of continuing anonymity; and

6.6.2 it is kept under regular review and updated periodically, in order to take into account of technological and other changes which could increase the ability of third parties to identify participants.

UK Biobank / June 2021