biobank uk
Improving the health of future generations

Summary de-identification protocol

1. Introduction

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

1.1.1 This note ***does not*** specifically address the steps which UK Biobank adopts when it releases participant data to third parties for the purposes of linkage to health records, outcomes validation and other purposes relating to improving the quality of the Participant Data. For example, in order to be able to link to a Participant's death / cancer / HES / GP records, UK Biobank has to release sufficient data about the Participant to the linkage counterparty. These arrangements are covered under formal agreements between UK Biobank and the relevant counterparty, which impose explicit contractual obligations on both UK Biobank and the counterparty;

1.1.2 Further, this note ***does not*** specifically address the simple fact of being identified as a Participant in UK Biobank. Indeed this is a fact that many Participants choose to volunteer about themselves. Nevertheless, UK Biobank will not of itself release or confirm the identity of any Participant;

1.1.3 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 *in conjunction with* inadvertently placing certain health-related information about the Participant into the public domain.

1.2 To illustrate: the fact that *Mr Jones* is a participant in UK Biobank is significant but not of itself overtly critical. Nevertheless, releasing information about the state of Mr Jones health to researchers in such a way that Mr Jones is identified and can be referenced directly to his phenotypical information would be highly critical. The purpose of this note is to set out the steps that UK Biobank will take to ensure that this does not happen.

1.3 This note sets out the nature of the Participant Data which UK Biobank holds on Participants, UK Biobank's legal obligations, practical consideration about identification and finally the protocol itself.

2. Background

2.1 Participant Data includes the following:

2.1.1 The phenotypical data about the Participant collected during the assessment visit (including identifiable data such as name and address);

2.1.2 The health-records data to which UK Biobank has secured linkage, namely data from the death and cancer registries, the HES data and in due course primary care data;

2.1.3 The data derived by researchers using the Participant Data and data derived from biochemical assays and other analyses[1] conducted on Participant samples; and

2.1.4 Further phenotypical data collected from re-assessment or enhancement visits or from re-contact studies.

2.2 The Participant Data is held by UK Biobank in reverse anonymised form. This is achieved by (a) removing the Participant's name and encrypting their NHS number (b) using a number of different identifiers for each Participant (each identifier is linked by a series of codes to discrete categories of data) and (c) these data categories are stored at appropriate levels (with appropriate user-access controls) within the security firewalls in line with the sensitivity of the data.

2.3 As such, the reverse anonymisation is a multi-layered process, with a range of reverse anonymisation codes being used: with each code used to link to particular categories of data. It follows that there are algorithms for reversing the anonymity: UK Biobank uses these when, for example, it contacts its Participants. Access to these algorithms is only available to those individuals within UK Biobank whose responsibilities specifically require it.

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 the undertakings provided to Participants during the consent process. These 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" only applies to identifiable individuals; and

3.2.2 UK Biobank owes certain duties to participants under the UK Data Protection Act not to release data to researchers in such a way that participants can be identified[2].

3.3 UK Biobank ensures that researchers comply with these obligations relating to Participant Data through the mechanism of its Material Transfer Agreement (*MTA).* In this regard, Participant Data will be only released to researchers (whose application has been accepted by UK Biobank) who agree to enter into UK Biobank's MTA, which sets out in detail what researchers are entitled to do with Participant Data. The MTA also contains specific provisions prohibiting researchers from trying to pro-actively identify or contact Participants.

4. Practical considerations

4.1 To recap, as UK Biobank has an obligation to provide Participant Data to researchers in a manner which (a) preserves the anonymity of its Participants and (b) does not enable

---

1 Which in due course will include genetic sequence data (whether that is in the form of a full or partial sequence or a selection of common SNPs.

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

Participants to be inadvertently identified, the principal steps that UK Biobank adopts are to remove the Participant's name and encrypt his/her NHS number.

4.2 However, there remain certain items of information within the Participant Data, which have varying degrees of potential to identify a Participant (either alone or in combination). These include: date of birth; gender; name of GP; ethnicity; postcode; event dates, such as admission to a particular hospital; and unedited free text fields in linkage health-record data relating to the Participant.

4.3 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 and date of birth) can serve to increase the risk of identification.

4.4 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 would have to possess (a) another comparable genetic sequence of the Participant from a source which identified the Participant and 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. As technology improves this situation may change and UK Biobank will keep it under review.

5. UK Biobank's protocol

5.1 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.

5.2 Typically, in terms of the Participant Data provided to researchers:

5.2.1 UK Biobank will, for example, only provide (a) grid references to varying levels of accuracy, e.g. 1km, 5km *rather than* a specific post code and (b) the Participant's month and year of birth *rather than* the d.o.b;

5.2.2 UK Biobank will not provide unedited free text fields nor the name of the Participant's GP; and

5.2.3 UK Biobank will generally provide gender, ethnicity and event dates (but not the healthcare location) about Participants.

5.3 It may be the case that researchers will request, for *bona fide* research purposes, more detailed information such as detailed grid references[3], which could increase the risk of Participant identification. In which case UK Biobank will adopt the following approach (using grid references as an example):

5.3.1 the specific grid reference would be released but no other information about the Participant would be released to the researcher; and

[3] This would be relevant if the researcher was trying to determine a specific location-based exposure (such as proximity to power lines).

5.3.2 UK Biobank would then perform the linkage of the derived data provided by the researcher to the Participant Data and return the appropriate data to the researcher.

5.4 UK Biobank would make two final caveats about this protocol:

5.4.1 it is not a fail-safe guarantee of anonymity;

5.4.2 it will probably need to be updated periodically, in order to take into account of technological and other changes which could increase the ability of third parties to identify participants.