UK Biobank / guidance note / de-duplication of common participants

1. The purpose of this guidance note, which relates to the identification and analysis of common participants, is to set out the approved mechanisms by which a researcher can de-duplicate common or related participants.

2. A researcher undertaking a combined analysis of UK Biobank and 3rd party data should avoid the double counting of an individual who is a participant (or a close relative thereof) in both UK Biobank and the 3rd party resource and UK Biobank. Such de-duplication is self-evidently good research practice (and standard in GWAS analysis) in order to remove unnecessary bias.

3. This de-duplication involves the matching of common participants – in a manner which mitigates any risk of re-identification – and then removal of the common participant such that double counting is avoided. This can be done in one of the following ways:

   - on request to (and by) UK Biobank (particularly if there is no genetic data involved in the de-duplication);
   - using a UK Biobank tool which is freely available in the Research Analysis Platform for this purpose; or
   - by the researcher themselves using an approved genotype-based tool.

4. In the latter case of the researcher conducting the matching using genetic data, this must be done as a *black box analysis* such that the output is solely a list of study participant identifiers to be removed. There are certain standard genotype-based tools that can do this, for example, PLINK or KING. In this way, the researcher is able to select their own parameters for the de-duplication, such as the degree of relatedness.

5. UK Biobank would remind researchers that it is not permitted to use the matched genotype data for the purpose of re-identifying (or seeking to re-identify) any participant.