Reporting small numbers in results in research outputs[1] using UK Biobank data

1. Background to the guidance

UK Biobank takes its participant's confidentiality and potential re-identification risk seriously. UK Biobank has reviewed best practice approaches on the publication of small numbers within a research publication (and related outputs) with a view to providing guidance to researchers on this topic.

This guidance provides researchers with information on how to report small numbers in any research output generated from the resource. It takes into account the published guidance from:

- the Office of National Statistics: https://www.ons.gov.uk/methodology/methodologytopicsandstatisticalconcepts/disclosurecontrol/policyonprotectingconfidentialityintablesofbirthanddeathstatistics; and
- NHS England: https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/hospital-episode-statistics/change-to-disclosure-control-methodology-for-hes-and-ecds-from-september-2018.

This guidance has been developed with the following factors in mind:

- The data provided by UK Biobank to researchers is always de-identified (with direct and indirect identifiers removed) in accordance with UK Biobank's de-identification protocol: https://www.ukbiobank.ac.uk/media/3kmh2t1z/de-identification-protocol-v-2-1-29-06-22.pdf
- UK Biobank considers it is appropriate to restrict the amount of individual-level participant data that can be made publicly available, as UK Biobank considers that it is not desirable to publish extensive details of individual participant phenotype (even if the participant remains de-identified);
- UK Biobank enables the use of certain web-based tools (such as phenotype browsers) to be deployed, so that a third party can have relevant information and a clear visualization and understanding (particularly in terms of cross tabulations) about what is in the UK Biobank resource, but not so that substantive clinical research can be conducted (without making an application to UK Biobank).

[1] Research outputs include (but are not limited to) peer-reviewed publications, pre-prints, twitter posts, blog posts, and web-based results browsers.

2. General guidance

UK Biobank makes the following general recommendations to researchers:

- To always seek to limit the amount of individual participant level data which is published; and
- To consider how, when constructing tables of results for publication, can be done in a way which mitigates any potential re-identification risk.

3. Presenting uncommon or unusual variables

UK Biobank considers that there are preferred ways of presenting or aggregating certain uncommon data points, in a manner which mitigates the risk of re-identification of UK Biobank participants. For example:

- The need to avoid the combination of two or more *uncommon components of phenotype,* for example presenting a rare disease code, such as G10 (Motor neurone disease), with a sparsely populated post code (on which further detail below), such as KW17 (Orkney), as this may increase potential re-identification risk particularly when taken in conjunction with information available on social media posts;
- Whereas the combination of only one uncommon component of phenotype in conjunction with a common component, such as presenting a rare disease code in combination with gender (even taking into account social media posts) has a much lower potential for re-identification.

4. **More specific suggestions** (which will be reviewed and updated as necessary)

The following are more specific suggestions based upon researcher enquiries:

- Web-based browsers, using summary statistics or results (e.g., GWAS-PheWAS browser), developed by a researcher, should contain a minimum number of 100 participants within a cell to be reported;
- Tables and/or results of any UK Biobank data (including linked healthcare record data), there should have a minimum number of 5 reported participants within a cell;
- For sample size queries by prospective researchers to UK Biobank (e.g., cross-tabulations of particular conditions or data fields) a minimum number of 5 participants will be reported;
- The use of scanned images of individual participants in research outputs is acceptable, as long as any participant level information included is minimised (e.g., age bands should be used where appropriate). The use of any UK Biobank image requires an appropriate *credit © UK Biobank, made available under licence.*

5. Geographical data

The Showcase database within UK Biobank contains a number of geographical fields which are shown below. These are available to researchers, although:

- the more highly resolved geographical data require a specific application to UK Biobank; and
- there are restrictions on the availability of census data taken in tandem with location co-ordinates (as the intersection between the two can be very specific).

Group	Precision	Restriction
Home location coordinates	1km	Not for release with home location census/admin areas. If <50 participants within the grid cell, will be released as a 10km grid
	100m	Requires a special justification to UK Biobank and a phased application
Home location census/admin areas	Lower layer Super Output Area, Middle layer Super Output Area, or Local Authority District (and Scottish equivalents) – see Resource 1406 for details on census geographies	Not for release with home location co-ordinates
	Output Area – see Resource 1406 for details on census geographies	Requires a special justification to UK Biobank and a phased application
Environmental derived variables	Various; sometimes derived from postcodes, 100m or 1m coordinates	None

In the event that further advice or clarification is required, please contact UK Biobank (access@ukbiobank.ac.uk).

UK Biobank / September 2023