Genes & Health

Policy – open disclosure of phenotypes and variants with counts and small numbers

This document

v2.5, author: David van Heel

approved by Genes & Health Exec: 6 Mar 2024

**open disclosure*: as used here, means disclosure of results outside the Genes & Health TRE, and no longer within Genes & Health control. This could include dissemination within a research applicants group, university, company, or website or publication.

Several approved researcher groups have requested data out of counts of phenotypes for variants/genotypes. This would be very useful to them for internal discussions and comparisons. In some cases, these data might be made generally available (e.g. via website or downloadable).

Some other studies completely suppress data on phenotype counts<=5 (e.g. Finngen). However Genes & Health offers scientific insights in rare diseases/phenotypes associated with unusual genotypes. It would be a missed opportunity to suppress this data. The Information Commissioners Office offers guidance (inference control) here https://ico.org.uk/media/1061/anonymisation-code.pdf. For Genes & Health the presence (count 1-5) of a phenotype versus absence (count 0) is usually more important than the actual count (between 1 and 5).

Researchers would output actual counts for >5, and for counts <=5 replace with a “1to5” count label. This would be checked at the time of data out request using our standard process.

We would only allow such requests for curated phenotypes, or null phenotypes (see below).

1. Open data, curated phenotypes

   The data would be a list of variants and genotypes, along with a list with counts of observed phenotypes. These phenotypes will **only be made from our curated_phenotypes** dataset (specifically the data in the TRE in /genesandhealth/library-red/phenotypes_curated/), which are usually aggregated across datasets and are either **coarse definitions** (e.g. 3 digit ICD10 “type 2 diabetes” rather than “pancreatic insufficiency diabetes with R leg ulcer”); or are a codelist name “type 2 diabetes” comprised of multiple specific ICD10/OPCS/SNOMED codes in a codelist, **such that going back to the individual code and raw data is not possible**. We would not allow phenotypes from raw_phenotypes or questionnaire data included. Phenotype frequency and count data (without genotype) is already openly available. Requests to export such data from the TRE would be subject to standard data out review.

If the number of participants with a variant(s) of interest is 2-5, curated_phenotypes can be exported for this group as per the standard data out review (described as either ‘0’ or ‘1-5’ in the report) process, similar to the scenario above.

If only one participant, need to follow the 3. Specific request for Exec review, below.

Example:

variant 16_124566_G_A, genotype AA: Crohn disease n=20; lung tuberculosis n=1to5; type 2 diabetes n=50.

2. Open dataset, specific detailed review of an individual genotype – nothing remarkable found

A brief summary that a review of all curated_phenotypes, raw_phenotypes, quantitative phenotypes, questionnaire information, and/or a live NHS health record review by Genes & Health staff (if available) **where nothing remarkable was found could be reported** as such, with description of phenotypes assessed but without numbers or counts.

Requests to export such data from the TRE would be subject to standard data out review.

Example:

variant 16_124566_G_A, genotype AA: all raw_phenotypes datasets and live Barts Health NHS Trust and Summary Care Record data were reviewed for 20 volunteers with this genotype. Common diseases such as diabetes were observed, but nothing clinically remarkable that might be considered related to genotype was noted. Inference control (replacing <=5 counts with ‘1to5’) would need to be applied.

3. Specific request for Exec review

More specific requests for more detailed data or for some unusual trait (not covered by the above) or quantitative trait data, and if there is a strong scientific case to make such information open (e.g. a manuscript for scientific publication) then a request could be made to the Genes & Health Executive for review.

Example:

We identified 1 individual with a homozygous knockout genotype (chrXX_12345_AA, pCys123STOP) in the NEWDRUG1 gene. The researchers at PharmaCo consider this information relevant (as a possible safety signal) to new drug development, and would like to include this in a publication. This female of British Bangladeshi ethnicity aged between 20-30 years had recurrent severe atypical fungal infections, and atypical/non-traumatic long bone fractures. These are very rare unusual conditions, and in the opinion of the clinical reviewer may be due to the genotype. Where possible data would be made less identifiable (e.g. decade age range, not actual age) using inference control.

Inference Control

For small numbers, to reduce the risk of identification, we will apply inference control (as advised by the Information Commissioners Office)
https://ico.org.uk/media/1061/anonymisation-code.pdf
Specifically, counts between 1 – 5 have the individual number replaced by the text “1to5”. We will also follow other recommendations in the Information Commissioners Office document.