Overfitting in polygenic risk score analyses: Exploring the impact of sample overlap and first degree relatives

Sarah Medland

QIMR Berghofer Medical Research Institute, Australia

QIMR Berghofer
Medical Research Institute

How important is independence with Biobank size samples?

Polygenic Risk scores (PRS) provide a quantitative measure of the cumulative genetic risk or vulnerability that an individual possesses for a trait.

To avoid overfitting the discovery and the target sample need to be independent.

There are perceptions that this may not matter with biobank type discovery samples when the overlap is very small.

Impact of relatedness across the discovery and target samples is usually ignored.

Methods

To examine this GWAS were conducted for a continuous (height) and a binary trait (results not shown)

- ~340,000 individuals were extracted from the UK Biobank (app. 25331)
- European Ancestry & Unrelated (< 3rd degree relatedness)
- Age, Sex and 10 PCs included as covariates
- A set of 35,000 individuals held out

- Discovery GWAS were clumped and PRS were calculated using 2,000, 5,000 or 10,000 individuals randomly drawn from the hold-out sample (of 35,000)
  - 1,000 replicates,
  - 4 PRS thresholds (results shown for p <= 0.0001)
  - Age, Sex and 10 PCs included as covariates
- To examine overfitting the target samples were spiked with
  - 5, 10, 50, 100 or 200 overlapping individuals
  - 5, 10, 50, 100 or 200 1st degree relatives

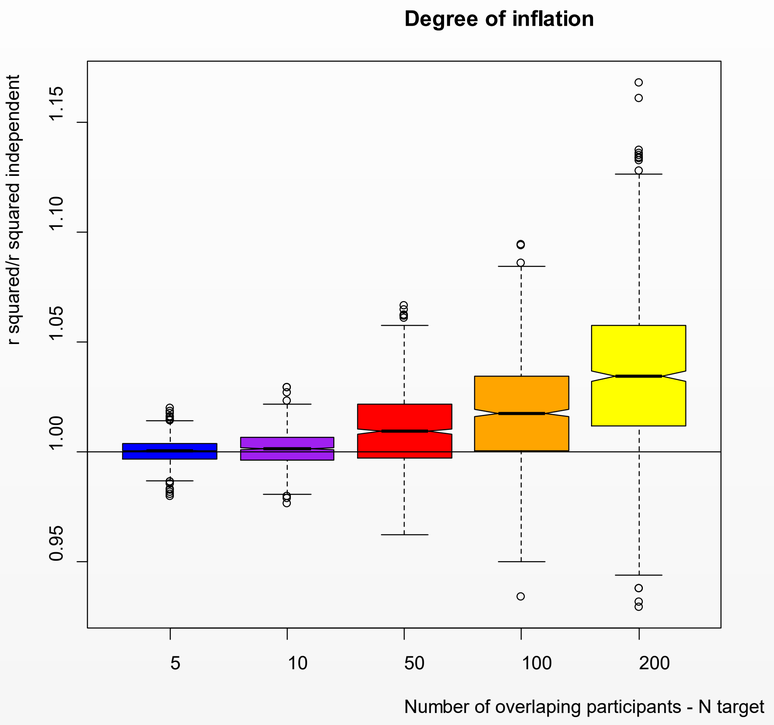
Impact of non-independent samples

As expected there is bias in the estimate of variance explained and the p values

Pattern of results the same across all Ns

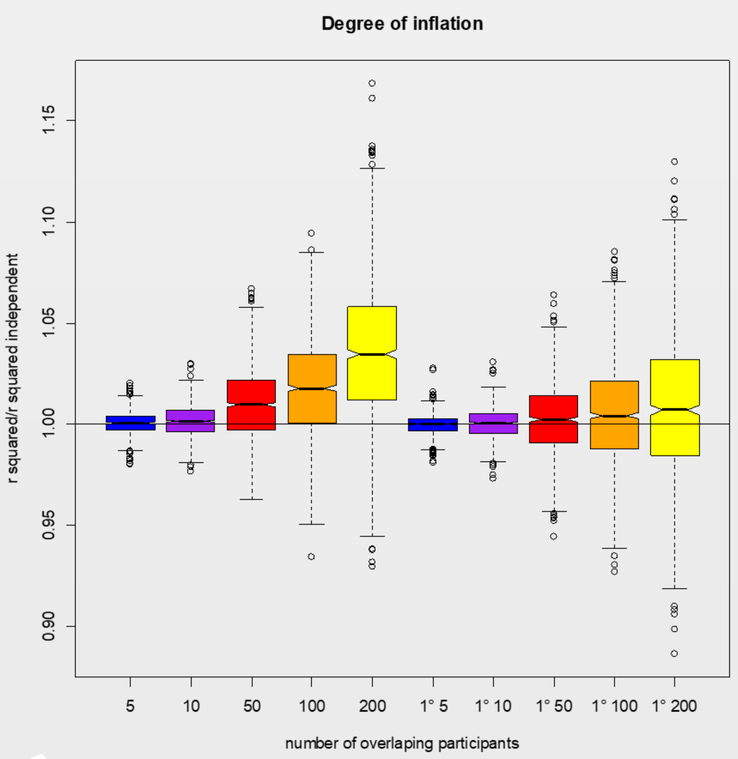
Degree of inflation is a function of the % overlap in the target sample

Inflation also present in binary phenotypes. Even if the overlap is limited to only controls or only cases



Impact of First Degree Relatives

Inflation present, proportional to the h^2 and the extent of overlap in the target sample (% of N, results shown for N=2,000).



How to Identify non-independence?

Option 1: Use the Homer et al[1] or Visscher and Hill[2] method.
Problem: However, many cohorts do not provide true MAF, these analyses typically violate data access, not clear how well this really works with a realistic meta-analysis where N (cohorts and participants) vary by SNP.

Option 2: Use LDScore and examine intercepts
Problem: Many target samples are too small to run LDscore and many PRS applications are cross-trait

What are the solutions if you find non-independence

LOO: Option a leave-one-out GWAS to create the PRS.

Checksums: If both groups have raw data access collaborate & exchange checksums

- Make list of common non-ambiguous SNPs passing QC in discovery and target
- Make *n* SNP set lists each with *m* SNPs
- Export hardcall data from each SNP set (1 line per person but no IDs)
- Parse the data obtaining a checksum for each line of data
- Exchange and look at % of identical checksums
- Problem: This won't find overlap of relatives.

Conclusions

Non-independence of discovery and target samples results in overfitting even with biobank level discovery sample sizes. Overlap of first degree relatives will also result in overfitting.

References:

[1] Homer N, Szelinger S, Redman M, Duggan D, Tembe W, et al. (2008) Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. PLoS Genet 4: e1000167

[2] Visscher PM, Hill WG (2009) The Limits of Individual Identification from Sample Allele Frequencies: Theory and Statistical Analysis. PLOS Genetics 5(10): e1000628.