Empirical evaluation of analytic validity of polygenic scores

Lin T[1], Zeng J[1], Gordon SD[2], Wallace LA[1], Ziser L[1], Shah S[1], Nolte IM[3], Snieder H[3], Lifelines Cohort Study, James PA[4], Martin NG[2], Visscher PM[1,5], Lee E[6], Yengo L[1], Henders AK[1], Wray NR[1,7]

1. Institute for Molecular Bioscience, The University of Queensland, Brisbane, Australia
2. QIMR Berghofer Medical Research Institute, Brisbane, Australia
3. Department of Epidemiology, University of Groningen, University Medical Center Groningen, The Netherlands
4. Peter MacCallum Cancer Centre, Melbourne, Australia
5. Nuffield Department of Population Health, University of Oxford, United Kingdom
6. Preciselee, Sydney, NSW 2065, Australia
7. Department of Psychiatry, University of Oxford, United Kingdom

Background: Polygenic scores (PGSs) are weighted sum scores of trait-associated alleles from up to millions of SNPs. As PGS research pivots to translation into health care settings a key issue for laboratories providing PGS is demonstration of analytical validity of PGS.

Methods. We report data from 6 individuals who have been genotyped multiple times using the same and different technologies, as part of standard experimental design for quality control purposed in two research settings over many studies and over many years. Using this opportunistic design of technical variability, we provide an empirical evaluation of technical reproducibility of PGS from 115 traits of different genetic architectures.

Results. Given a predefined set of SNP weights variability in PGS can reflect only SNP missingness or incorrect genotype call. We find very high reproducibility of SNP genotypes. In particular, the technical reproducibility of PGS generated from the same array technology and processed through the same quality control and imputation pipeline is very high. However, impact of missing SNPs varies between traits depending on the SNPs weight for a trait.

Conclusions. Objective approaches to evaluate analytical reproducibility of PGS which can use imputed genotypes has been lacking. Here, we provide empirical data and an analysis framework which can be used by PGS providers to support understanding of analytical reproducibility and robustness.