Understanding Society Polygenic Scores

USER GUIDE

Version 2, *June 2025*

Authors: Anna Dearman, Yanchun Bao and Meena Kumari

Institute for Social and Economic Research
University of Essex
Colchester
Essex

Understanding Society

Economic and Social Research Council

Contents

1. Introduction

Understanding Society is a longitudinal household panel survey that started in 2009 and attempts to interview the core sample members at approximately one year intervals, as long as they are living in the UK. Each sets of interviews are referred to as waves. After the 2nd and 3rd waves sub-samples of the Study consented to nurse visits, blood and DNA extraction. To learn more about the survey please visit the About the Study webpage and the Main Survey User Guide. To learn more about the nurse visits and the health and biomarker data collected please visit the Health and biomarkers webpage.

A polygenic score (PGS) is a continuous variable that reflects an individual's propensity towards a given trait, as calculated by comparing their genetic profile to the summary statistics generated from a genome-wide association study (GWAS) of the trait. Traits are wide-ranging and can include disease status, behaviours, and blood levels of biomolecules, among many others. In the case of disease status, polygenic scores are sometimes termed "polygenic risk scores" (PRSs) as they reflect the risk, or liability, of having the disease.

For a given trait, the GWAS summary statistics comprise a long list of single-nucleotide polymorphisms (SNPs), or points on the genome which have two possible variants (alleles), which have been tested for association with the trait of interest. For each SNP, one of the alleles, often referred to as the "effect allele", has an effect size and p-value reported in the summary statistics. An individual has 0, 1 or 2 copies of the effect allele per SNP. PGSs are sums - usually weighted by effect size - of the effect alleles that an individual possesses. A biological glossary can be found on the Understanding Society biomarker topic page.

Each PGS that is made available in *Understanding* Society has a separate list of all SNPs included in the calculation. This user guide includes the GWAS paper and details of the methods used to generate the PGS in *Understanding Society*. PGS calculation methods vary in terms of the software used, the exclusion or inclusion of SNPs based on different clumping methods and p-value thresholds, weighted vs unweighted approaches, etc.

2. Polygenic Score of Body Mass Index in Understanding Society

The polygenic score (PGS) of body mass index (BMI) is based on the genome-wide association study (GWAS) of BMI by Yengo et al (2018) [1] and the PGS calculated by Hughes et al (2019) [2]. The imputed genetic dataset from *Understanding Society* was used to calculate individual-level PGSs. The following formula was used:

$$PGS = \sum_j^m SNP_j\, \beta_j,\ \ j = 1, 2, \dots, m, \qquad (1)$$

where SNP represents single-nucleotide polymorphism. SNP_j is the jth SNP of m SNPs that are significantly associated with BMI under GWAS p-value threshold ($p < 5 \times 10^{-8}$). β_j is the effect size of jth SNP according to the GWAS [1].

The working steps are:

1. SNPs were first excluded on the basis of:
   - minor allele frequency <0.01
   - Hardy-Weinberg p<1x10^{-4}
   - call rate <0.95
   - imputation information score <0.8

2. SNPs were identified as genome-wide significant from published GWAS information [1] ($p < 5 \times 10^{-8}$). Then the significant SNPs were clumped using *clump_data()* of *TwoSampleMR* package in MRBase at clump.kb=10000, clump.r2=0.001, resulting in 480 SNPs used in the score [3 4]. Please refer to the file "UKHLS_BMI_PGS_SNPs.csv" for details of 480 SNPs.

3. The PGS was then calculated for each individual based on the formula (1) using R code.

See references for Body Mass Index PGS in the Reference section.

3. Polygenic Score of Testosterone in Understanding Society

The polygenic score (PGS) of total testosterone (TT) is based on the genome-wide association study (GWAS) of TT by Ohlsson et al (2011) [1] and the PGS calculated by Hughes and Kumari (2019) [2]. The PGS is based on only three SNPs, two of which were directly genotyped in *Understanding Society* and one which was only available in the imputed dataset. The following formula was used:

$$PGS = \sum_j^m SNP_j\, \beta_j,\ \ j = 1, 2, \ldots, m, \qquad (1)$$

where SNP represents single-nucleotide polymorphism. SNP_j is the jth SNP of m SNPs that are significantly associated with BMI under GWAS p-value threshold ($p < 5 \times 10^{-8}$). β_j is the effect size of jth SNP according to the GWAS [1].

The working steps are:

1. No SNPs were excluded. The GWAS reported only three independent, significantly associated SNPs [1] and all three were used to calculate the PGS in *Understanding Society*. Please refer to the file "UKHLS_TT_PGS_SNPs.csv" for details of the three SNPs.

2. The PGS was then calculated for each individual based on the formula (1) using STATA code.

See references for Testosterone PGS in the Reference section.

4. Data and Analysis Guidance

Users may find the following tutorial useful as it provides 'theoretical background and hands-on experience' for researchers new to the field - A tutorial on conducting genome-wide association studies: Quality control and statistical analysis – PubMed (nih.gov)[1].

4.1 List of files and variables

File:	Description:
xpolygen_ns.dta	Polygenic data
BMI_pgs_snps.csv	List of Body Mass Index SNPs
Testosterone_pgs_snps.csv	List of testosterone SNPs

Variable:	Label:
pidp	cross-wave person identifier (public release)
pid	personal identifier (BHPS cohort)
b_hidp	household identifier (public release)
b_pno	person number
b_splitnum	split number, 1=first, 2=second
c_hidp	household identifier (public release)
c_pno	person number
c_splitnum	split number, 1=first, 2=second
wave	health assessment wave
hhorig	sample origin BHPS or Understanding Society
pgs_BMI_exwt_dv	BMI polygenic score
pgs_testosterone_exwt_dv	testosterone polygenic score
pca1	genetic principal component PCA1
pca2	genetic principal component PCA2
pca3	genetic principal component PCA3
pca4	genetic principal component PCA4
pca5	genetic principal component PCA5
pca6	genetic principal component PCA6
pca7	genetic principal component PCA7
pca8	genetic principal component PCA8
pca9	genetic principal component PCA9
pca10	genetic principal component PCA10

[1] Marees AT, de Kluiver H, Stringer S, Vorspan F, Curis E, Marie-Claire C, Derks EM. A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. Int J Methods Psychiatr Res. 2018 Jun;27(2):e1608. doi: 10.1002/mpr.1608. Epub 2018 Feb 27. PMID: 29484742; PMCID: PMC6001694.

5. Discussion and Recommendations

PGSs generated in *Understanding Society* can be used as explanatory variables in a range of analyses, including as an instrumental variable (IV) in Mendelian Randomization (MR) studies to investigate the causal effect of the PGS trait with health/social outcomes – for examples using the PGS of body mass index (BMI) please see [1, 2]. MR entails a special set of assumptions; for more information about MR see [3-6].

Any findings from analyses that use PGSs as explanatory variables require careful interpretation. Users are strongly advised to read the corresponding GWAS paper in the PGS reference sections. It is important to note aspects of the GWAS cohort(s) which may be relevant to the measurement of the trait, such as age and health, and it is especially crucial to note the ancestral population(s), as single-ancestry PGSs have limited portability across different populations due to factors such as differences in allele frequencies and patterns of linkage disequilibrium. It is also important to note any measures of the PGS's explained variance and predictive power included in the GWAS paper. Finally, genetic associations can occur due to environmental confounding, especially in traits that are more behavioural, social or psychological. More discussion on the application of PGS in social research can be found in [7-12].

6. Data Access

These Data are released through the UK Data Service. The Special Licence (SL) or safeguarded version, SN7587, can be found here: https://beta.ukdataservice.ac.uk/datacatalogue/studies/study?id=7587.

These datasets are categorized as Special Licence and therefore have more restrictive access conditions, details of which can be found in the "Access data" tab at the above link. In addition, access to these Special Licence datasets requires users to complete forms justifying why the data is needed. Further details on the application process can be found here: https://ukdataservice.ac.uk/find-data/access-conditions/

7. Citation

7.1 Citing the data

If you use *Understanding Society* data you must cite every study that you use.

The citation for the data can be found at https://www.understandingsociety.ac.uk/documentation/citation

All works which use or refer to these materials should acknowledge these sources by means of bibliographic citation. To ensure that such source attributions are captured for bibliographic indexes, citations must appear in footnotes or in the reference section of publications.

7.2 Citing this User Guide

When citing this User Guide, you can use the citation of this particular version quoted below. Note that where an online version is available on the Understanding Society website it is always the most up to date.

Institute for Social and Economic Research. (2025), *Understanding Society:* Polygenic Scores*, User Guide, Version 2, June 2025*, Colchester: University of Essex.

8. Online resources and documentation

Information about Understanding Society health, biomarkers, genetics and epigenetic (DNA methylation) data can be found on the Understanding Society website at https://www.understandingsociety.ac.uk/documentation/health-assessment/. Specifically, you can find information on:

- How to access the data, user guides
- An interactive variable search facility
- Questionnaires
- Fieldwork documents
- Generic FAQs
- User guides. This section includes different user guides starting with the “Nurse Assessment User Guide” which describes the health and biomarker data collected by nurses. In addition to this user guide, this section includes the “Biomarker User Guide and Glossary”, “Proteomics User Guide”, “Epigenetic Ageing Algorithms (clocks) User Guide”. This section also includes a “Biological data glossary”.

9. References

[1] Hughes A, Bao Y, Smart M, Kumari M. Body mass index, earnings and partnership: genetic instrumental variable analysis in two nationally representative UK samples. bioRxiv 2019. DOI: 10.1101/608588

[2] Howe LD, Kanayalal R, Beaumont R, Davies AR, Frayling TM, Harrison S, et al. Effects of body mass index on relationship status, social contact, and socioeconomic position: Mendelian Randomization study in UK Biobank. 2019:524488

[3] Burgess S, Thompson SG. Use of allele scores as instrumental variables for Mendelian randomization. International Journal of Epidemiology. 2013;42(4):1134-44.

[4] Davies NM, Holmes MV, Davey Smith G. Reading Mendelian randomisation studies: a guide, glossary, and checklist for clinicians. BMJ. 2018;362.

[5] Burgess S, Bowden J, Fall T, Ingelsson E, Thompson SG. Sensitivity Analyses for Robust Causal Inference from Mendelian Randomization Analyses with Multiple Genetic Variants. Epidemiology. 2017;28(1):30-42.

[6] Bowden J, Del Greco M F, Minelli C, Davey Smith G, Sheehan N, Thompson J. A framework for the investigation of pleiotropy in two-sample summary data Mendelian randomization. Statistics in Medicine. 2017;36(11):1783-802.

[7] Morris TT, Davies NM, Hemani G, Smith GD. Population phenomena inflate genetic associations of complex social traits. Science advances 6 (16), eaay0328

[8] Barth D, Papageorge N W, Thom K, Genetic Ability, Wealth, and Financial Decision-Making. IZA discussion paper series, 2017. IZA DP No. 10567.

[9] Beauchamp, J P. Genetic evidence for natural selection in humans in the contemporary United States. PNAS, 2016. 113 (**28**): 7774-7779.

[10] Belsky D W, Moffitt T E, Corcoran D L, et al. The genetics of success: how single nucleotide polymorphisms associated with educational attainment relate to life-course development. Psychological Science, 2016. 27 (**7**): 957-972.

[11] Munafò, M. , Davies, N. M., & Davey Smith, G. (2020). Can genetics reveal the causes and consequences of educational attainment? Journal of the Royal Statistical Society: Series A, 183(2), 681-688. https://doi.org/10.1111/rssa.12543

[12] Abdellaoui, A., Verweij, K.J.H. Dissecting polygenic signals from genome-wide association studies on human behaviour. Nat Hum Behav 5, 686–694 (2021). https://doi.org/10.1038/s41562-021-01110-y

References for Body Mass Index PGS:

[1] Yengo L, Sidorenko J, Kemper KE, et.al. Meta-analysis of genome-wide association studies for height and body mass index in ~700000 individuals of European ancestry. Hum Mol Genet. 2018 Oct 15;27(20):3641-3649.

[2] Hughes A, Bao Y, Smart M, Kumari M. Body mass index, earnings and partnership: genetic instrumental variable analysis in two nationally representative UK samples. bioRxiv 2019 608588; doi: https://doi.org/10.1101/608588

[3] Marees A T, De Kluiver H, Stringer S, et.al. A tutorial on conducting genome-wide association studies: quality control and statistical analysis. Int J Methods Psychiatr Res. 2018; **27**: e1608.

[4] Hemani G, Zheng J, Elsworth B, Wade KH, Baird D, Haberland V, Laurin C, Burgess S, Bowden J, Langdon R, Tan VY, Yarmolinsky J, Shihab HA, Timpson NJ, Evans DM, Relton C, Martin RM, Davey Smith G, Gaunt TR, Haycock PC, The MR-Base Collaboration. The MR-Base platform supports systematic causal inference across the human phenome. eLife 2018;7:e34408.

References for Testosterone PGS:

[1] Ohlsson C, Wallaschofski H, Lunetta KL, Stolk L, Perry JR, Koster A, Petersen AK, Eriksson J, Lehtimäki T, Huhtaniemi IT, Hammond GL, Maggio M, Coviello AD; EMAS Study Group, Ferrucci L, Heier M, Hofman A, Holliday KL, Jansson JO, Kähönen M, Karasik D, Karlsson MK, Kiel DP, Liu Y, Ljunggren O, Lorentzon M, Lyytikäinen LP, Meitinger T, Mellström D, Melzer D, Miljkovic I, Nauck M, Nilsson M, Penninx B, Pye SR, Vasan RS, Reincke M, Rivadeneira F, Tajar A, Teumer A, Uitterlinden AG, Ulloor J, Viikari J, Völker U, Völzke H, Wichmann HE, Wu TS, Zhuang WV, Ziv E, Wu FC, Raitakari O, Eriksson A, Bidlingmaier M, Harris TB, Murray A, de Jong FH, Murabito JM, Bhasin S, Vandenput L, Haring R. Genetic determinants of serum testosterone concentrations in men. PLoS Genet. 2011 Oct;7(10):e1002313. doi: 10.1371/journal.pgen.1002313

[2] Hughes A, Kumari M. Testosterone, risk, and socioeconomic position in British men: Exploring causal directionality. Social Science & Medicine. 2019; 220: 129-140. doi: 10.1016/j.socscimed.2018.11.004