Polygenic indices in four national longitudinal cohorts

User Guide (Version 1)

August 2025

CENTRE FOR LONGITUDINAL STUDIES

UKRI Economic and Social Research Council

Contact

Data queries: help@ukdataservice.ac.uk

Questions and feedback about this user guide: clsdata@ucl.ac.uk.

Authors

Tim T Morris, Gemma Shireby, Liam Wright, Aida Sanchez-Galvez, Georg Otto, David Bann.

How to cite this guide

Morris TT, Shireby G, Wright L, Sanchez-Galvez A, Otto G, Bann D. (2025). *Polygenic indices in four national longitudinal cohorts*. London: UCL Centre for Longitudinal Studies.

Data citation and acknowledgement

You should cite the data and acknowledge CLS following the guidance from cls.ucl.ac.uk/data-access-training/citing-our-data/. All outputs using CLS genetic data should cite this user guide and the following article:

Shireby G, Morris TT, Wong A, Chaturvedi N, Ploubidis GB, Fitzsimmons E, Goodman A, Sanchez-Galvez A, Davies NM, Wright L, Bann D. *Data Resource Profile: Genomic Data in Multiple British Birth Cohorts (1946-2001)-Health, Social, and Environmental Data from Birth to Old Age*. medRxiv. 2024:2024-11.

Centre for Longitudinal Studies

Centre for Longitudinal Studies (CLS)

UCL Social Research Institute

University College London

20 Bedford Way, London WC1H 0AL

www.cls.ucl.ac.uk

The UCL Centre for Longitudinal Studies (CLS) is an Economic and Social Research Council (ESRC) Resource Centre. It is home to a unique series of UK national cohort

studies. It is part of the UCL Social Research Institute, based at the IOE, UCL's Faculty of Education and Society.

This document is available in alternative formats. Please email the Centre for Longitudinal Studies at clsdata@ucl.ac.uk.

Acknowledgements

The CLS cohorts are only possible due to the commitment and enthusiasm of their participants; their time and contribution are gratefully acknowledged. We thank colleagues in the Centre for Longitudinal Studies Research Data Management Team, Survey, Cohort Maintenance, Administrative, and Communications teams; and colleagues at the University of Bristol for their work on biosample assay and storage.

Contents

1. Introduction

This document describes the process by which polygenic indexes (PGIs) for 44 traits have been constructed in four UK longitudinal cohort studies that follow large nationally representative groups of people since birth. Three of these are national birth cohort studies initiated in 1958 (National Child Development Study, 1958c) [1] 1970 (British Cohort Study, 1970c) [2,3] and 2000 (Millennium Cohort Study, 2001c) [4] and one is a cohort born in 1989-90 followed up from adolescence (Next Steps, 1989c) [5].

The PGIs have been developed using a consistent methodology that has been applied to harmonised genetic data across each cohort, enabling researchers to engage in consistent cross-cohort analysis for using derived genetic measures for the first time. All PGIs have been derived from large scale Genome-wide Association Studies (GWAS) with publicly available summary statistics. Through this approach we hope to enable and encourage wider use of the genetic data collected in these studies. We also provide high level guidance on the use and interpretation of PGIs.

The PGIs were also developed in a consistent manner in a birth cohort born in 1946 (MRC National Survey of Health and Development, 1946c) [6], which can be obtained by separate application to the Unit for Lifelong Health and Ageing at UCL.

2. Project background

2.1 Rationale

Biosamples such as blood and saliva are now commonly collected from participants in cohort and longitudinal studies, from which genome-wide genetic information is derived via genotyping. However, these studies are often highly selective or not nationally representative, limiting the generalisations that can be drawn from their data. The availability of genetic data from large scale nationally representative datasets with known sampling distributions such as those included in this data release therefore offers unique opportunities for research.

The increasing availability of genetic data has enabled a wide range of analyses that have improved our understanding of human biology. The overwhelming majority of human traits analysed in genetic studies have been shown to implicate multiple genetic variants that each explain a very small amount of statistical variance in the trait. This pattern of traits being associated with multiple variants is termed polygenicity, and the effects of these variants can be combined into a single score that represents the known genetic predisposition to a trait known as a polygenic index (also “polygenic scores” or “polygenic risk scores”; we opted to use the term ‘index’ to avoid value judgments that ‘score’ can imply) [7]. Polygenic indexes have become an increasingly popular tool by which researchers can better understand human behaviour and use methods that compliment ‘traditional’ observational or survey-based approaches which do not directly measure genetic factors [8].

2.2 The human genome

The human genome is composed of Deoxyribonucleic acid (DNA), which is organized into structures called chromosomes and stored within our cells. Most people possess 23 pairs of chromosomes, with one set inherited from each biological parent. DNA itself is formed from two intertwined strands, structured as a double helix, and made up of units known as nucleotides. These nucleotides differ

based on one of four chemical bases: adenine (A), cytosine (C), guanine (G), and thymine (T).

While the human genome includes roughly 3.2 billion nucleotide base pairs, the overwhelming majority of these are the same across all individuals. The points of the genome at which nucleotide base pairs differ – the sites of genetic variation - are known as single nucleotide polymorphisms, or SNPs (pronounced “snips”). At these points, individuals may have different chemical bases (e.g., an A instead of a G), and the versions present are referred to as alleles. Because chromosomes come in pairs, each person can carry zero, one, or two copies of a particular allele at any given SNP.

2.3 Genome-wide Association Studies (GWAS)

To understand the construction of PGI it is first necessary to understand GWAS. GWAS are large-scale projects that combine data from collections of individuals – often from multiple cohort and longitudinal studies - who have provided genetic and survey or direct assessment-based (phenotypic) data. GWAS sizes are often on the order of millions of participants, with the largest sample size to date being 5.4m [9]. For each trait a regression is run at each SNP in a discovery phase where the trait (e.g., height) is the dependent variable and the count of effect alleles (i.e. 0, 1, 2) is the independent variable. Given the number of SNPs in the human genome that are measured by genotyping arrays (on the order of millions), stringent p-value thresholds are applied by GWAS to account for multiple testing, typically at $p<5x10^{-8}$. Individuals can be pooled into a single sample or these GWAS effects meta-analysed across samples. For each SNP, GWAS produce an estimate of the average effect size for each additional effect allele across the sample.

To reduce risks of overfitting within the sample and to test generalisability, GWAS perform out-of-sample validation on cohort or longitudinal studies that are not included in the discovery phase. This step helps to ensure that the effect sizes identified in the discovery phase are consistent in an independent sample. GWAS will usually report the predictive performance of the combined estimates in the

independent as the amount of variation that they explain in the trait (R^2) over independent variables such as age and sex.

GWAS effect sizes are commonly published as publicly available GWAS summary statistics that are available through online resources such as the NHGRI-EBI GWAS Catalogue and the IEU OpenGWAS project.

2.4 Polygenicity

GWAS have demonstrated that most human traits are implicated by a wide range of genetic variants i.e., they are polygenic. That is, traits for which there can be considered to "be a single genetic variant for" are the exception rather than the norm. Examples of single-gene conditions include cystic fibrosis and Huntington's disease. Polygenic traits are characterised by many (often thousands) of SNPs that each have a very small effect size. For example, the largest GWAS to date implicated 12,111 independent SNPs associated with height [9].

2.5 Polygenic indexes

PGIs aggregate GWAS estimates across all measured SNPs to provide a single estimate of an individual's genetic predisposition towards the trait under study. As such, SNPs can be considered as the building blocks of PGIs. It is important to note that the genetic predisposition represented by PGI is known inasmuch as it has been estimated accurately and reliably from a GWAS; not all SNPs are included in the GWAS or the GWAS estimates for a given SNP are inaccurate, then the genetic predisposition represented by the PGI will be lower than the true genetic predisposition.

3. Derivation of polygenic indexes

3.1 PGI theory

The central model underlying the construction of PGIs is as follows:

$$PGI_i = \sum_{j=1}^{M} \beta_j G_{ij}$$

Where PGI_i is the estimated PGI for individual i calculated as the sum over M SNPs, where each SNP j is weighted by its effect size β_j (estimated from GWAS) and multiplied by the number of risk-increasing alleles G_{ij} (0, 1, or 2) that individual i carries at SNP j. As such, every PGI contains data from two sources: the external input GWAS (known as the 'base data'), and the genotypes of individuals within the sample (known as the 'target data'). By summarising the count of alleles at multiple SNPs, PGI approximate continuous variables that are normally distributed in the population.

3.2 Linkage Disequilibrium

To ensure the predictive validity of PGIs, it is important to account for Linkage Disequilibrium (LD): the non-random association of genetic variants that are located near each other in the genome and tend to be inherited together. LD can lead to overlapping genetic signals being counted multiple times, which may inflate the influence of certain genomic regions in the PGI [10]. The derivation of PGIs is an active area of research and multiple methods now exist to both generate PGIs and account for LD [11]. Existing guides and reviews provide overviews of these methods [11–13], as well as discussing the relevance of PGIs to social science [8,14].

The current version of the CLS PGI repository uses a clumping and thresholding (C+T) approach to account for LD when generating PGIs. Clumping selects one representative SNP from a set of highly correlated SNPs (based on a specified LD threshold) to minimise effects due to LD. Thresholding refers to the selection of

SNPs from a GWAS to be used in a PGI, for example all SNPs whose estimates meet a threshold of GWAS significance ($p<5x10^{-8}$).

3.3 PGI transformation

To ease interpretation, PGIs are often standardised in the sample to standard deviation / z scores (mean of 0 and SD of 1):

$$PGI_i^{std} = \frac{PGI_i - \overline{PGI}}{\sigma_{PGI}}$$

The PGIs calculated by the CLS PGI pipeline are presented in raw format; it is recommended that analysts standardise these prior to use.

3.4 CLS PGI pipeline

The full pipeline code by which PGIs were derived is provided on the CLS GitHub. Briefly, the below steps were taken by the pipeline and applied to the genetic data that had been quality controlled as outlined by [15].

1. Preparation of the trait list from which PGIs were to be derived. This involved downloading GWAS summary statistics and formatting the trait list in such a way that it could be read by the pipeline.
2. Download reference files for GRCh38 Genome Reference Consortium Human Build 38 and subset chromosome 3 for checking genome builds of the GWAS summary statistics.
3. Creation of a harmonised SNP list across all cohorts to be used in cross-cohort harmonised PGIs (harmonised SNP n=6,702,716) to optimise cross-cohort comparability. Note that in this first release of the CLS PGI repository PGIs are only provided for cohort participants who are more genetically similar to 1000 Genomes Phase 3 samples labelled as European and self-report as being of a white ethnic background.
4. Genome build conversion to GRCh38 and quality control of the GWAS summary statistics for each trait.

5. Generation of PGIs using PRSice2. Multiple PGIs were generated for each trait, as follows:
    a. Using the maximum SNP lists for each cohort independently and using the harmonised SNP lists across all cohorts.
    b. Across five p-value thresholds (p<5x10-8; p<5x10-5. p<0.05; p<0.01; p=1).

3.5 Sample sizes

Table 1 below shows the maximum and minimum SNP coverage within and across all cohorts restricting to variants with valid rsID's.

Table 1. SNP coverage in the cohorts (valid rsID's)

Cohort	N (maximum available SNP coverage within each cohort)	N (harmonised SNP coverage across all cohorts)
NCDS	7,545,089	6,703,052
BCS	8,094,234	6,703,052
Next Steps	8,084,092	6,703,052
Millennium Cohort Study	8,412,240	6,703,052

3.6 List of PGI traits

Table 2 below lists the PGIs created for the cohort studies. We chose a range of traits that are applicable to a broad range of disciplinary specialities and have been studied in large-scale GWAS.

Traits included in the CLS PGI repository will increase in future as further GWAS studies become available. For the most up to date information on GWAS included please see the GitHub page at https://cls-genetics.github.io/.

PGI's are available for all cohort members (and in the MCS, the cohort members biological parents) who provided biosamples that passed all quality control steps.

Table 2. Polygenic index traits

Domain	Trait	Reference
Anthropometrics	Birth weight	16
	Body fat distribution	17
	Body Mass Index (childhood)	18
	Body Mass Index (adulthood)	19
	Grip strength	20
	Height	19
	Waist circumference	21
Brain structure and cognition	Alzheimer's disease	22
	Cognition	23
	Hippocampal volume	24
	Parkinson's disease	25
Health behaviours	Substance abuse	26
	Age at initiation of smoking	27
	Alcoholic drinks per week	27
	Cigarettes per day	27
	Diet	28
Mental health	Anxiety	29
	ADHD	30
	Autism spectrum disorder	31
	Bipolar disorder	32
	Depressive symptoms	33
	Externalising problems	34
	Major depressive disorder	35
	Schizophrenia	36
Personality	Agreeableness	37
	Conscientiousness	37
	Extraversion	37
	Openness to experience	37
	Neuroticism	37
Physical health	Age at menopause	38
	Asthma	39
	Blood pressure	40
	Coronary artery disease	41

Domain	Trait	Reference
	C-reactive protein	42
	Fasting glucose	43
	HbA1c	44
	Hypertension	45
	Rheumatoid arthritis	46
	Type 1 Diabetes	47
	Type 2 Diabetes	48
Social outcomes	Education	49
	Household Income	50
	Human Longevity	51
	Parental Lifespan	52

4. Guidance on interpreting PGIs

4.1 Interpreting PGIs

We encourage users to read previous guides and tutorials which describe in detail how PGIs could be used and how they should and should not be interpreted [12–14]. Users should also read and familiarise themselves with the CLS Resource Profile paper which discusses the genetic data underlying the PGI and potential uses of cohort data in genetically informed designs [15]. Users of MCS data are also advised to refer to [53] which contains detailed information on collection of the biological data, protocols, and response patterns.

PGI are probabilistic summary indicators of genetic predisposition towards traits and as such users should avoid any overextrapolation of results or genetic determinism when using the PGIs. Because of LD and genotyping coverage PGI will typically capture both causal and non-causal variants. Furthermore, PGIs for complex and biologically distal traits have been shown to be inflated by population level demographic and familial factors, including population stratification, which can bias GWAS estimates and inflate PGI associations [54–56].

Users are strongly encouraged to review the GWAS paper(s) used to construct the PGI(s) they use. PGIs should be interpreted in the context of the underlying GWAS, specifically its sample composition and size, trait (phenotype) definition and measurement, sample ancestry and the statistical methods used. Users should also check whether the GWAS paper used contains CLS cohorts: this is very unlikely to be the case for more recently genotyped cohorts (1970c, 1989c, 2001c) but is more likely for 1958c which was part of the Wellcome Trust Case Control Consortium [57].

The PGIs have been released as raw scores to enable users to apply multiple imputation and/or Inverse Probability Weighting to their projects. We recommend that users transform the PGIs to z scores prior to analyses for ease of interpretation. For further information and detailed guidance on the use of multiple imputation and Inverse Probability Weighting in the CLS cohorts, please see the Handling Missing Data section on the CLS website.

4.2 Ancestry and genetic similarity

Many existing GWAS (from which our PGI are derived) have restricted their study samples to groups that are relatively genetically homogenous. These samples are often described in terms of "genetic ancestry", which is commonly assigned using arbitrary cutoffs of genetic similarity to other groups of individuals from a reference database such as the 1000 Genomes Phase 3. That is, genetic ancestry labels are heavily simplified statistical modelling constructs drawn from external sources. An individual labelled as "European genetic ancestry" is simply defined as such by virtue of having a genotype that is more similar to individuals in a reference dataset who are also labelled as "European". For an excellent discussion of genetic similarity and ancestry, users are recommended to read [58], and for detailed guidance on the use of population descriptors in genetic research, to read [59]. While this approach is somewhat blunt and reductive, it can be necessary given the difficulty of accurately modelling real-world complexity and the need to simplify reality to meet modelling assumptions (e.g., of relatively homogenous samples).

This first release of the CLS PGI repository contains PGIs only for cohort participants (and in the case of the MCS, the cohort participant's biological parents) who are genetically similar to 1000 Genomes Phase 3 samples labelled as European as defined using an elastic net model and <4 standard deviations of the mean of the first principal component of genetic similarity. The decision to only include these individuals in the first release was because of inconsistencies in the performance of PGIs across broader population groups and the potential subsequent introduction of bias to studies. Table A1 in the appendix displays the number of individuals whose samples passed QC in the CLS cohorts that have been included/excluded in the CLS PGI repository.

4.3 Suggested wording for outputs using the CLS PGI repository

Given that the CLS PGI repository currently only contains PGIs for a subset of individuals across the cohorts, researchers are encouraged to note this, the potential

limitations, and provide a brief explanation. Below we provide an example that researchers may wish to build upon.

*Our study was limited to individuals who were genetically similar to European samples in 1000 Genomes Phase 3 as defined using an elastic net model (**delete as appropriate**) in order to minimise sample heterogeneity and limit bias due to differential PGI performance across study individuals / to aid comparability with previous studies / given the use of less diverse cohorts / since comparable polygenic indexes were not available for all individuals. This limits the generalisability of our findings to the broader population. As genetic knowledge improves across the genetic similarity spectrum, future research should extend these findings to the entire population.*

5. Description of the research datasets

5.1 Licensing and data access

The PGI datasets have been processed by CLS and supplied to the UK Data Service (UKDS). All data users need to be registered with the UK Data Service and to sign the UKDS End User Licence. Details of how to do this are available at the UKDS website.

The PGI datasets have been pseudonymized and are available from the UKDS as special safeguarded data, which are subject to the UKDS Special Licence. The UKDS Special Licence application form requires information on the goals of the research project, and should address any risks related to the ethical, sensitive and/or disclosive nature of the research topic, explaining how these will be managed or mitigated. For guidance on how to manage these risks, please refer to Section 6.6 (*Research with disclosive data or in socially controversial areas*) of the CLS Data Access Framework. Once the form has been reviewed by UK Data Service and approved by the CLS Data Access Committee the data will be available to download.

5.2 List of datasets

Datasets are in a wide format i.e., a single row per participant. The datasets available in this release are listed in Table 3 below.

Table 3: List of available datasets

Name of the dataset	Content summary
CLS_PGI_v1_NCDS	Polygenic indices in NCDS
CLS_PGI_v1_BCS	Polygenic indices in BCS70
CLS_PGI_v1_MCS	Polygenic indices in the Millennium Cohort Study
CLS_PGI_v1_NS	Polygenic indices in Next Steps

5.3 Identifiers

For all cohorts the data are identified with the same research IDs used for the rest of cohort data available at the UK Data Service (e.g., NCDSID for NCDS). This enables the data to be merged with one another datasets held on UKDS. Note that the MCS uses family identifiers (MCSID) and the two individual person identifiers (CNUM00/PNUM00). Further information on merging is available in the CLS data management GitHub repository and the MCS Data Handling Guide.

5.4 Variable names and labels

The variable naming convention is consistent across the datasets, designed such that the specific PGI that a researcher is using can always be traced back to source in future versions of the CLS PGI repository. The variable names contain the following elements:

- Cohort
- If PGI is from harmonised SNP list
- Trait name
- p value threshold
- PGI repository version

For example, the NCDS harmonised PGI for height at a p value threshold of p<5x10-8 would be named ***ncds_hmz_eight_p5e08_v1***.

The variable labelling convention is consistent across the datasets, consisting of the following elements:

- If PGI is from harmonised SNP list
- Trait name
- GWAS author and year
- p value threshold

For example, the NCDS harmonised PGI for height at a p value threshold of p<5x10-8 would be labelled as ***Harmonised PGI for height from Yengo et al (2018) at p5e08***.

5.5 Missing values

Missing values are present for all individuals who did not provide biosamples or whose samples did not pass quality control checks.

6. Reproducibility

To maximise transparency and reproducibility, the code pipeline to generate all PGIs in this data release from the source genetic data is available on the CLS Data GitHub. The pipeline can be cloned and applied by users with access to the full genomewide genetic data to generate further PGIs that are not part of the repository in a consistent manner. The pipeline contains a detailed readme and requires minimal user input to run.

7. References

1. Power, C. & Elliott, J. Cohort profile: 1958 British birth cohort (National Child Development Study). *Int J Epidemiol* **35**, (2006).
2. Elliott, J. & Shepherd, P. Cohort profile: 1970 British Birth Cohort (BCS70). *Int J Epidemiol* **35**, (2006).
3. Sullivan, A., Brown, M., Hamer, M. & Ploubidis, G. B. Cohort Profile Update: The 1970 British Cohort Study (BCS70). *Int J Epidemiol* **52**, (2023).
4. Connelly, R. & Platt, L. Cohort profile: UK Millennium Cohort Study (mcs). *Int J Epidemiol* **43**, (2014).
5. Wu, A. F.-W. *et al.* Cohort Profile: Next Steps—the longitudinal study of people in England born in 1989–90. *Int J Epidemiol* **53**, dyae152 (2024).
6. Wadsworth, M., Kuh, D., Richards, M. & Hardy, R. Cohort profile: The 1946 National Birth Cohort (MRC National Survey of Health and Development). *Int J Epidemiol* **35**, (2006).
7. Becker, J. *et al.* Resource profile and user guide of the Polygenic Index Repository. *Nat Hum Behav* **5**, (2021).
8. Morris, T. T. *et al.* Implications of the genomic revolution for education research and policy. *Br Educ Res J* (2022) doi:10.1002/berj.3784.
9. Yengo, L. *et al.* A saturated map of common genetic variants associated with human height. *Nature* **610**, (2022).
10. Dudbridge, F. Polygenic Epidemiology. *Genet Epidemiol* **40**, 268–272 (2016).
11. Ni, G. *et al.* A Comparison of Ten Polygenic Score Methods for Psychiatric Disorders Applied Across Multiple Cohorts. *Biol Psychiatry* **90**, (2021).
12. Choi, S. W., Mak, T. S. H. & O'Reilly, P. F. Tutorial: a guide to performing polygenic risk score analyses. *Nature Protocols* vol. 15 Preprint at https://doi.org/10.1038/s41596-020-0353-1 (2020).

13. Allegrini, A. G., Baldwin, J. R., Barkhuizen, W. & Pingault, J. B. Research Review: A guide to computing and implementing polygenic scores in developmental research. *Journal of Child Psychology and Psychiatry and Allied Disciplines* vol. 63 Preprint at https://doi.org/10.1111/jcpp.13611 (2022).

14. Burt, C. H. Polygenic Indices (aka Polygenic Scores) in Social Science: A Guide for Interpretation and Evaluation. *Sociol Methodol* **54**, (2024).

15. Shireby, G. *et al.* Data Resource Profile: Genomic Data in Multiple British Birth Cohorts (1946-2001) - Health, Social, and Environmental Data from Birth to Old Age. *medRxiv* (2024).

16. Warrington, N. M. *et al.* Maternal and fetal genetic effects on birth weight and their relevance to cardio-metabolic risk factors. *Nat Genet* **51**, (2019).

17. Pulit, S. L. *et al.* Meta-Analysis of genome-wide association studies for body fat distribution in 694 649 individuals of European ancestry. *Hum Mol Genet* **28**, (2019).

18. Vogelezang, S. *et al.* Novel loci for childhood body mass index and shared heritability with adult cardiometabolic traits. *PLoS Genet* **16**, (2020).

19. Yengo, L. *et al.* Meta-analysis of genome-wide association studies for height and body mass index in ~700 000 individuals of European ancestry. *Hum Mol Genet* **27**, (2018).

20. Jones, G. *et al.* Genome-wide meta-analysis of muscle weakness identifies 15 susceptibility loci in older men and women. *Nat Commun* **12**, (2021).

21. Christakoudi, S., Evangelou, E., Riboli, E. & Tsilidis, K. K. GWAS of allometric body-shape indices in UK Biobank identifies loci suggesting associations with morphogenesis, organogenesis, adrenal cell renewal and cancer. *Sci Rep* **11**, (2021).

22. Bellenguez, C. *et al.* New insights into the genetic etiology of Alzheimer's disease and related dementias. *Nat Genet* **54**, (2022).

23. Savage, J. E. *et al.* Genome-wide association meta-analysis in 269,867 individuals identifies new genetic and functional links to intelligence. *Nat Genet* **50**, (2018).

24. Liu, N. *et al.* Cross-ancestry genome-wide association meta-analyses of hippocampal and subfield volumes. *Nat Genet* **55**, (2023).

25. Nalls, M. A. *et al.* Identification of novel risk loci, causal insights, and heritable risk for Parkinson's disease: a meta-analysis of genome-wide association studies. *Lancet Neurol* **18**, (2019).

26. Hatoum, A. S. *et al.* Multivariate genome-wide association meta-analysis of over 1 million subjects identifies loci underlying multiple substance use disorders. *Nature Mental Health* **1**, (2023).

27. Liu, M. *et al.* Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use. *Nature Genetics* vol. 51 Preprint at https://doi.org/10.1038/s41588-018-0307-5 (2019).

28. Cole, J. B., Florez, J. C. & Hirschhorn, J. N. Comprehensive genomic analysis of dietary habits in UK Biobank identifies hundreds of genetic associations. *Nat Commun* **11**, (2020).

29. Forstner, A. J. *et al.* Genome-wide association study of panic disorder reveals genetic overlap with neuroticism and depression. *Mol Psychiatry* **26**, (2021).

30. Demontis, D. *et al.* Genome-wide analyses of ADHD identify 27 risk loci, refine the genetic architecture and implicate several cognitive domains. *Nat Genet* **55**, (2023).

31. Grove, J. *et al.* Identification of common genetic risk variants for autism spectrum disorder. *Nat Genet* **51**, (2019).

32. Mullins, N. *et al.* Genome-wide association study of more than 40,000 bipolar disorder cases provides new insights into the underlying biology. *Nat Genet* **53**, (2021).

33. Baselmans, B. M. L. *et al.* Multivariate genome-wide analyses of the well-being spectrum. *Nat Genet* **51**, (2019).

34. Karlsson Linnér, R. *et al.* Multivariate analysis of 1.5 million people identifies genetic associations with traits related to self-regulation and addiction. *Nat Neurosci* **24**, (2021).

35. Howard, D. M. *et al.* Genome-wide meta-analysis of depression identifies 102 independent variants and highlights the importance of the prefrontal brain regions. *Nat Neurosci* **22**, (2019).

36. Trubetskoy, V. *et al.* Mapping genomic loci implicates genes and synaptic biology in schizophrenia. *Nature* **604**, (2022).

37. Gupta, P. *et al.* A genome-wide investigation into the underlying genetic architecture of personality traits and overlap with psychopathology. *Nat Hum Behav* (2024) doi:10.1038/s41562-024-01951-3.

38. Ruth, K. S. *et al.* Genetic insights into biological mechanisms governing human ovarian ageing. *Nature* **596**, (2021).

39. Han, Y. *et al.* Genome-wide analysis highlights contribution of immune system pathways to the genetic architecture of asthma. *Nat Commun* **11**, (2020).

40. Keaton, J. M. *et al.* Genome-wide analysis in over 1 million individuals of European ancestry yields improved polygenic risk scores for blood pressure traits. *Nat Genet* **56**, 778–791 (2024).

41. Aragam, K. G. *et al.* Discovery and systematic characterization of risk variants and genes for coronary artery disease in over a million participants. *Nat Genet* **54**, (2022).

42. Koskeridis, F. *et al.* Pleiotropic genetic architecture and novel loci for C-reactive protein levels. *Nat Commun* **13**, (2022).

43. Downie, C. G. *et al.* Multi-ethnic GWAS and fine-mapping of glycaemic traits identify novel loci in the PAGE Study. *Diabetologia* **65**, (2022).

44. Sinnott-Armstrong, N. *et al.* Genetics of 35 blood and urine biomarkers in the UK Biobank. *Nat Genet* **53**, (2021).

45. Bi, W., Fritsche, L. G., Mukherjee, B., Kim, S. & Lee, S. A Fast and Accurate Method for Genome-Wide Time-to-Event Data Analysis and Its Application to UK Biobank. *Am J Hum Genet* **107**, (2020).

46. Ishigaki, K. *et al.* Multi-ancestry genome-wide association analyses identify novel genetic mechanisms in rheumatoid arthritis. *Nat Genet* **54**, (2022).

47. Chiou, J. *et al.* Interpreting type 1 diabetes risk with genetics and single-cell epigenomics. *Nature* **594**, (2021).

48. Suzuki, K. *et al.* Genetic drivers of heterogeneity in type 2 diabetes pathophysiology. *Nature* **627**, (2024).

49. Okbay, A. *et al.* Polygenic prediction of educational attainment within and between families from genome-wide association analyses in 3 million individuals. *Nat Genet* **54**, (2022).

50. Hill, W. D. *et al.* Genome-wide analysis identifies molecular systems and 149 genetic loci associated with income. *Nat Commun* (2019) doi:10.1038/s41467-019-13585-5.

51. Pilling, L. C. *et al.* Human longevity: 25 genetic loci associated in 389,166 UK biobank participants. *Aging* **9**, (2017).

52. Timmers, P. R. H. J. *et al.* Genomics of 1 million parent lifespans implicates novel pathways and common diseases and distinguishes survival chances. *Elife* **8**, (2019).

53. Fitzsimons, E. *et al.* Collection of genetic data at scale for a nationally representative population: the UK Millennium Cohort Study. *Longit Life Course Stud* **13**, (2022).

54. Morris, T. T., Davies, N. M., Hemani, G. & Smith, G. D. Population phenomena inflate genetic associations of complex social traits. *Sci Adv* **6**, (2020).

55. Howe, L. J. *et al.* Within-sibship genome-wide association analyses decrease bias in estimates of direct genetic effects. *Nat Genet* **54**, (2022).

56. Kong, A. *et al.* The nature of nurture: Effects of parental genotypes. *Science (1979)* **359**, 424–428 (2018).

57. Burton, P. R. *et al.* Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, (2007).

58. Coop, G. Genetic similarity versus genetic ancestry groups as sample descriptors in human genetics. *arXiv preprint arXiv:2207.11595* (2022).

59. National Academies of Sciences Engineering and Medicine. *Using Population Descriptors in Genetics and Genomics Research: A New Framework for an Evolving Field*. *Using Population Descriptors in Genetics and Genomics Research: A New Framework for an Evolving Field* (The National Academies Press, Washington DC, 2023). doi:10.17226/26902.

8. Appendix

Table A1: Samples included/excluded in the CLS PGI repository across the CLS studies.

Cohort	Individuals passed genetic QC	Individuals excluded from PGI repository*	Individuals included in PGI repository*
NCDS	6,396	72	6,396
BCS	5,598	237	5,361
Next Steps	1,568	296	1,272
Millennium Cohort Study	20,247	3,142	17,105

* Individuals were included in the CLS PGI repository if they were genetically similar to European samples in 1000 Genomes Phase 3 using an elastic net model, and <4 standard deviations from mean of the first principal component of population structure.