DOCUMENTATION REPORT

THE ENGLISH LONGITUDINAL STUDY OF AGEING (ELSA)

POLYGENIC SCORES

2022

Report prepared by
Dr Olesya Ajnakina, University College London
Professor Andrew Steptoe, University College London

Contact details: o.ajnakina@ucl.ac.uk
Department of Behavioural Science and Health
Institute of Epidemiology and Health Care
University College London
1-19 Torrington Place, London WC1E 7H

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

1.1. Overview

This document describes the construction of polygenic scores (PGSs) for a number of behavioural, emotional and health-related phenotypes in the English longitudinal study of ageing (ELSA) study. The methods employed for creating PGSs descried herein are those outlined by the Health and Retirement Study (HRS)[1]. This was done in order to harmonise the research across age-related longitudinal studies by adopting a consistent methodology for creating PGSs. By making these PGSs publicly available, it is hoped that they will facilitate wide use among the ELSA data users. PGSs for each phenotype are based on a single, replicated genome-wide association study (GWAS). These scores will be updated as sufficiently large GWAS are published for new phenotypes or as updated meta-analyses for existing phenotypes are released. This document describes the methodology employed in creating PGSs through quality control to construction of the PG scores and presents an overview of PGSs for these phenotypes in ELSA.

1.2. Rationale

Recent advances in technology have allowed the systematic hypothesis-free testing of genetic variants across the entire human genome for association with various traits measured on unrelated individuals [2-4]. However, for many complex genetic traits the well-powered GWASs did not identity individual markers that exceeded the Odds Ratio (OR) of more than 1.2, which is lower than initially anticipated (i.e., OR between 1.5-2) [5]. This in turn raised the question whether common variants in combination are of greater importance in the development of the phenotype than single variants with a large effect [3]. Indeed, many health and behavioural outcomes, such as smoking, obesity, Alzheimer's disease and schizophrenia, have been shown to be highly polygenic[2] implying that their genetic architecture consists of "many" genetic variants. Creating PGSs is a method that captures this signature. The methods that we employed for creating PGSs in the ELSA study will be described in more detail in Section 3.

1.3. The use of PGSs in scientific research

PG scores are usually constructed from a weighted sum of allelic count [3, 4, 6] and are presented as continuous scores. They are specific to each individual and represent an individual load for the common variants that are associated with a trait under study. PG scores are increasingly used to predict disease risks [6]. This is usually done through linear regression analyses where the PGS for a given trait is used a predictor for an outcome adjusting the analyses for various covariates, which usually age, gender and principal components to account for any ancestry differences in genetic structures that could bias results [7] (for more detail about principal components, please refer to Section 2.4.3.). Another popular way if using the PGSS is to derive a binary predictor from the continuous PGS, where the top 10% or 20% of the PGS is coded as "high risk" group and the remaining is coded as "low risk" group based on an individual loading for the common SNPs. In turn, genomic prediction of disease risks might have implications in designing more individualised preventive or screening strategies for patients [6]. For example, earlier screening for breast cancer may be warranted for those having a high genetic risk for the disease as measuring the PGS [8]. Furthermore, PGSs have been shown to be suitable for a number of scientific aims beyond the risk prediction including identification of shared aetiology among traits using such an analytical tool as GCTA

(Genome-wide Complex Trait Analysis)[9], testing for genome-wide G*E and G*G interactions[10], Mendelian Randomisation to infer causal relationships, and for patient stratification and sub-phenotyping[8, 11]. Thus, PGSs represent not only an individual genetic prediction of phenotypes but open possibilities for interrogating a wide range of hypotheses via association testing.

2. GWAS QUALITY CONTROL

2.1. Study participants

The English Longitudinal Study of Ageing (ELSA) is a large, multidisciplinary study of cohort of men and women living in England aged 50 or over and representative of the English population both in terms of socioeconomic profile and geographic region [12]. The study commenced in 2002 and the cohort was then followed-up every two years, with periodic refreshments to maintain the age profile. Since 2002 there have been 8 waves of data collection providing detailed information on health, well-being, and socioeconomic circumstances. Further, the ELSA study has been modelled on the US Health and Retirement Study (HRS) [13]. This was done to facilitate harmonisation with the HRS study and other ageing studies, and thus to promote international comparisons in the age-related outcomes across the population-based cohorts.

2.2. Consent and Administration Procedures

The ELSA participants were eligible for blood data collection if they had successfully completed the nurse visit and gave consent for blood samples to be taken. The respondents were not eligible to have a blood sample taken if they: 1) had a clotting or bleeding disorder, 2) ever had a fit or convulsion, 3) were taking anticoagulant drugs (such as Warfarin, Protamine or Acenocoumarol), or 4) were pregnant. If the ELSA participants were eligible to have a blood sample, nurses then determined whether they were eligible to fast. Those respondents who were determined to be eligible to fast, were instructed not eat, smoke, drink alcohol or do any vigorous exercise 30 minutes before giving the blood sample. The responders were exempted from fasting if they: 1) were aged 80 or over, 2) were diabetic and on treatment, or 3) were malnourished or otherwise unfit to fast (as judged by the nurse). All respondents could still drink water and take their medication as normal.

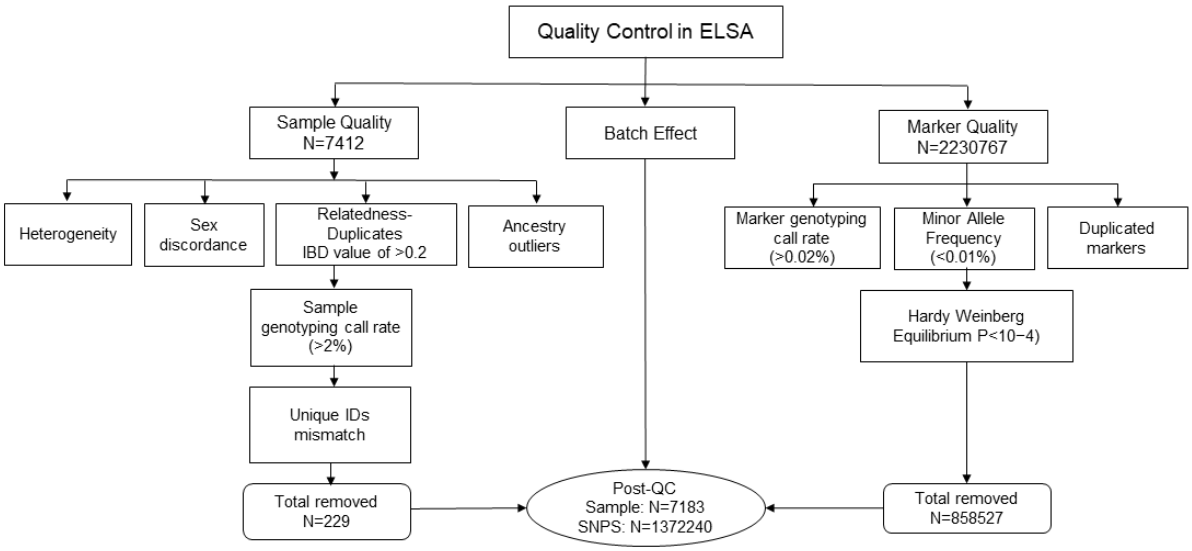
2.3. Genotyping Process

The genome-wide genotyping was performed at University College London (UCL) Genomics in 2013-2014. This involved genotyping of 7,597 ELSA participants of European ancestry using the Illumina HumanOmni2.5 BeadChips (HumanOmni2.5-4v1, HumanOmni2.5-8v1.3), which measures ~2.5 million markers that capture the genomic variation down to 2.5% minor allele frequency (MAF). Genotyping was performed in two batches. Allele frequencies were compared between the batches after filtering for 5% of missingness. The correlation was calculated between the batches for a number of chromosomes and exceeded 99%. After post-genotyping quality assurance, such as excluding ethnic outliers (self-reported) and duplicates, the GWAS data was available for total 7,412 ELSA participants and 2,230,767 SNPs.

2.4. GWAS Quality Control

Before the GWAS data was utilised for creating PGSs, a thorough quality control (QC) [14] at both individuals and single-nucleotide polymorphism (SNPs) levels was carried out using PLINK 1.9 [15]. The full QC procedure is depicted in **Figure 1**.

Figure 1. QC steps that were undertaken as part of quality control in ELSA



2.4.1. QC based on individual level

The samples for whom the recorded sex phenotype was inconsistent with genetic sex were removed. Duplicated samples and cryptic relatedness between each pair of participants was evaluated using pairwise genome-wide estimates of three coefficients corresponding to the probabilities of sharing 0, 1 or 2 alleles between two individuals that are identical by descent [16]. There are two methods for estimating the identical by descent (IBD) probabilities - method of moments and method of maximum likelihood. Both methods have been shown to give very similar results [17]; thus, we report results from method of moments implemented in PLINK 1.9 [15]. IBD were estimated using autosomal SNPs where IBD=1 highlights presence of duplicates or monozygotic twins, IBD=0.5 shows that first-degree relatives are present in the sample, IBD=0.25 and IBD=0.125 highlights presence of second-degree and third-degree relatives, respectively [18]. Owing to genotyping error, linkage disequilibrium (LD) and population structure, it is expected to observe some variations around these theoretical values. Therefore, it is normal to remove one individual from each pair with an IBD value of >0.2, which is halfway between the expected IBD for third- and second-degree relatives [14]. We identified individuals with an IBD value of >0.2 and excluded one of each pair at random.

2.4.2. QC based on SNP level

Heterozygosity refers to carrying of two different alleles of a specific SNP. Excessive heterozygosity may imply a sample contamination, while less heterozygosity than expected may imply inbreeding [14]. In the ELSA study, the checks for heterozygosity were performed on a set of SNPs which were non-(highly) correlated. To generate a list of non-(highly) correlated SNPs, we excluded four regions that are known to contain clusters of highly correlated SNPs. These were the Lactase Gene (LCT) (chromosome 6, 12578740 to 135837195 bp), human leukocyte antigen (HLA) (chromosome 2, 2550000 to 3350000 bp)

and two inversion regions located on 8p23.1 (chromosome 8, 81305000 to 1200000 bp) and 17q21.31 (chromosome 17, 40900000-45000000 bp) [19]. We then pruned the SNPs using the '10 5 0.1' parameters. These pruning parameters use a sliding window method that considers blocks of 10 SNPs and removes SNPs with r^2 >0.10 afterward shifting the window by 5 SNPs. Those individuals with extremely low or high heterozygosity score (>3 standard deviations from the mean) were removed. Further, the genotyped data with a call rate of <98% was removed. SNPs in sex chromosomes and SNPs with a minor allele frequency (MAF) of <0.01 were excluded. SNPs whose genotype distributions deviated significantly from the Hardy-Weinberg equilibrium (HWE) ($p<10^{-4}$) and with missingness <0.02 were also removed. Finally, to ensure a large overlap between the GWAS summary statistics (i.e., base file) and the ELSA (i.e., target) data, we have converted all present platform specific ids (i.e., kgps) to rsids. However, not all kgps were able to be successfully updated; those SNPs for which the kgps were not updated were removed.

2.4.3. Population structure

To investigate population structure, we use principal components analysis (PCA) [7] implemented in PLINK 1.9 [15]. We used the PCA approach with two aims; first, to identify those individuals who deviated from the ethnic population they self-reported to be (i.e., ethnic outliers), and second, to provide sample eigenvectors which will then be used for adjusting for possible population stratification in the association analyses [7, 20]. It has been shown that in PCA, the usefulness of certain principal components (PCs) may be limited by clusters of highly correlated SNPs at specific locations, such as the LCT, HLA, 8p23.1 and 17q21.31 [17, 19] in whole-genome arrays [19]. To address this pitfall, the SNPs that were used in PCA were selected by LD pruning from an initial pool consisting of all autosomal SNPs with a missing call rate <5% and MAF >5%. In addition, the 2q21 (LCT), HLA, 8p23, and 17q21.31 regions were excluded from this initial pool. The LD pruning process, using all unrelated ELSA participants selected 147,070 SNPs with all pairs having r^2 <0.1 in a sliding 10 Mb window. PCs were obtained using PLINK software; we retained the top 10 PCs to account for any ancestry differences in genetic structures that could bias results [7]. Initially, we performed PCA on all study subjects; however, the visual inspection of the PCs distribution highlighted the present of ancestral admixture in the 65 individuals. We removed these outliers and re-calculated PCs using the updated samples (**Supplementary Figure 1.**).

2.5. Summary of QC

After these QC steps 7223 (97.5% *n*=7412) individuals and 1,374,524 (61.5% of *n*=2230767 SNPs) directly genotyped SNPs remained for further analyses. The biggest proportion of the lost SNPs was due to MAF (34.1%); the remaining QC criteria led to loss of 0.1-2.2% of genotyped SNPs. The loss of ELSA participants was very minimal (between 0.07% and 1.0% of the total sample depending on the QC steps). Additionally, for 41 participants the ELSA Unique IDs was not available; these individuals were removed leaving the final sample of 7183.

Table 1. An overview of the summary of full QC procedure employed in the ELSA study and how many variants and/or participants were lost at each step.

Quality Control steps in ELSA		
Lost due to SNP-based QC	*n*	%
Missing SNPs (0.02)	41614	1.87
Autosomal SNPs	48578	2.18
MAF 0.01	759972	34.07
Update rsids	2284	0.10
HWE (0.0001)	6079	0.27
Total removed	*858527*	*38.49*
Total remaining	*1372240*	*61.51*
Lost due to Individual-based QC		
Missingness (0.02)	39	0.53
Heterogeneity	76	1.03
Sex discordance	5	0.07
Ancestry outliers	64	0.86
Relatedness/Duplicates	5	0.07
Unique IDs are not present	41	0.50
Total removed	*229*	*3.09*
Total remaining	*7183*	*96.91*

HWE, Hardy-Weinberg equilibrium; MAF, minor allele frequency; SNP, single nucleotide polymorphisms

3. POLYGENIC SCORE (PGS)

3.1. Overview of methodology

Polygenic scores (PGS) can be defined as a single value estimate of an individual's propensity to a phenotype, calculated as a sum of their genome-wide genotypes weighted by corresponding genotype effect sizes from GWAS summary statistics [3, 21]. Therefore, PGS analyses can be characterised by the two input data sets: 1) base (GWAS) data; these are summary statistics (e.g., betas, *p*-values) of genotype-phenotype associations at genetic variants (i.e., SNPs) in GWAS, and 2) target data; these are genotypes and phenotype(s) in individuals of the target sample (i.e., herein the ELSA data). A PGS is then calculated for each individual in the target sample following the formula outlined below:

$$\mathrm{PGS}^{i} = \sum_{j=1}^{j} W^{j} G^{ij}$$

where *i* is individual *i* (i=1 to N), *j* is SNP *j* (j=1 to J), *W* is the meta-analysis effect size for SNP *j* and *G* is the genotype, or the number of reference alleles (0, 1, or 2), for individual *i* at SNP *j*. The profile score is then evaluated through regression of the target sample phenotype on the PGS after accounting for other known covariates.

Because SNP effects are estimated with some uncertainty and not all SNPs influence the trait under study, PGSs are calculated at different pre-specified significance threshold of quality controlled and autosomal SNPs [3]. This in turn allows testing associations with the target trait for each threshold and thus optimising the prediction. Accordingly, we performed PGSs based on threshold of *p*-values of 0.001, 0.01, 0.05, 0.1, 0.3, and 1 employing methodology as originally described [1, 3]. Nonetheless, the HRS team examined four traits with large published and replicated GWASs (i.e., height, body mass index, educational attainment, and depression) demonstrating that PGSs that included all available SNPs either explained the most amount of variation in an outcome or were not significantly different from the PGSs that PGSs calculated at different *p*-value threshold. Thus, for reproducibility through rigor and transparency, they recommended that researchers include a PGS with all available SNPs as a reference, and provide substantial justification for using alternative methods[1]. Following this recommendation, we will make available the PGSs calculated for *p*-value threshold of 1. All the results related to PGSs reported herein will be based on this threshold.

Similarly to the HRS study[1], unless otherwise specified, if the beta/OR value from the GWAS summary statistics was negative (or the OR <1), the beta/OR measures were converted to positive values and the reference allele flipped to represent phenotype-increasing PGS. Moreover, we built the PGSs based on the directly genotyped data rather than imputed data. This decision was based on the previous research findings which highlighted that the PGSs built from the directly genotyped data had more predictive power [22] or did not differ significantly from the PGSs that were based on imputed data[1]. All analyses were restricted to individuals of European ancestry. These analyses were performed using PRSice [23] and PLINK 1.9 [15]. The PGSs that were made publicly available for ELSA users were not adjusted for any potential covariates when being constructed.

3.2. Sources for SNP weights

To incorporate externally valid SNP weights from GWASs, we performed a search of the literature to identify large GWAS meta-analysis studies related to the selected phenotype. Where possible, the meta-analyses that did not include ELSA in the discovery analysis were selected to be independent of our data. SNP weights were downloaded from the consortium webpages, requested from consortium authors, obtained from dbGap, or taken from published supplemental material. If ELSA was included in the analyses, we requested that the consortia to repeat the analysis with ELSA removed. All base SNP files from GWAS meta-analyses were converted to NCBI (National Center for Biotechnology Information) build 37 annotation for compatibility with ELSA SNP data.

3.3. RESULTS

3.3.1. PERSONALITY TYPES

3.3.1.1. Extraversion

The GWAS meta-analyses for Extraversion was conducted by the Genetics of Personality Consortium (GPC) [24]. The meta-analysis on Extroversion was performed on 63,030 subjects from 29 discovery cohorts. Sample sizes of the individual cohorts ranged from 177 to 7210 subjects. Extraversion scores were regressed on each SNP under an additive model, with sex and age included as covariates. Covariates such as ancestry PCs were added if deemed necessary for a particular cohort. Meta-analysis of GWA results did not yield genome-wide significant SNPs associated with Extraversion. The lowest p-value observed was $2.9x10^{-7}$ for a SNP located on chromosome 2. There were 74 SNPs with *P*-values $<1x10^{-5}$. The GWAS for Extraversion contained 6,941,603 SNPs; of these, 1,218,049 SNPs overlapped with the ELSA target data and were included in the PGS for Extraversion.

3.3.1.2. Agreeableness, Openness to Experience and Conscientiousness

The PGSs for Agreeableness, Openness to Experience and Conscientiousness were calculated based on the GWAS meta-analysis of Big Five personality traits [25]. TThis GWAS combined data from 10 studies, including 17375 individuals of European ancestry. In *silico* replication of the genome-wide significant SNPs was sought in five additional samples consisting of 3294 individuals. To compare results at the SNP level, ~2.5M common SNPs were imputed using the HapMap phase II CEU data as the reference sample. GWA analyses were conducted in each study independently using linear regression under an additive model and including sex and age as covariates. Two SNPs for Openness to Experience on chromosome 5q14.3 and one SNP for Conscientiousness on chromosome 18q21.1 passed the genome-wide significance level of $p<5\times10^{-8}$ in the discovery stage. No genome-wide significant results were found for Agreeableness

3.3.1.3. Neuroticism

PGS for Neuroticism was calculated based in the GWAS summary statistics that collated results from the Genetics of Personality Consortium (GPC) (n=63,661) and results from a new analysis of UKB data cohort (n=107,245) [26]. The meta-analysis yielded 11 lead SNPs, 2 of which tag inversion polymorphisms. In UKB, the phenotype measure was the respondent's score on a 12-item version of the Eysenck Personality Inventory Neuroticism scale. The GPC harmonised different neuroticism batteries. In the UKB, analyses controlled for the first 15 PCs, indicator variables for genotyping array, sex, indicator variables for age ranges, and sex-by-age interactions. Model adjustments for the 29 cohorts contributing to the GPC meta-analysis varied. The GWAS for Neuroticism contained 6,524,432 SNPs; of these, 1,191,041 SNPs overlapped with the ELSA target data and were included in the PGSs for Neuroticism.

Table 2. The summary statistics for PGSs for Extraversion, Agreeableness, Openness to Experience, Conscientiousness and Neuroticism in the ELSA study.

PGSs	Sample Size	Minimum	Maximum	Range	Median	Mean	SE (mean)
Agreeableness	7183	51706.9	52875.1	1168.2	52216.0	52216.2	1.71
Openness	7183	80286.5	80731.9	445.4	80527.4	80527.1	0.72
Conscientiousness	7183	62805.8	63155.0	350.0	62973.7	62974.2	0.56
Neuroticism	7183	5351.2	5459.7	108.4	5407.3	5407.1	0.20
Extraversion	7183	9386.5	9534.0	147.5	9462.6	9462.8	0.22

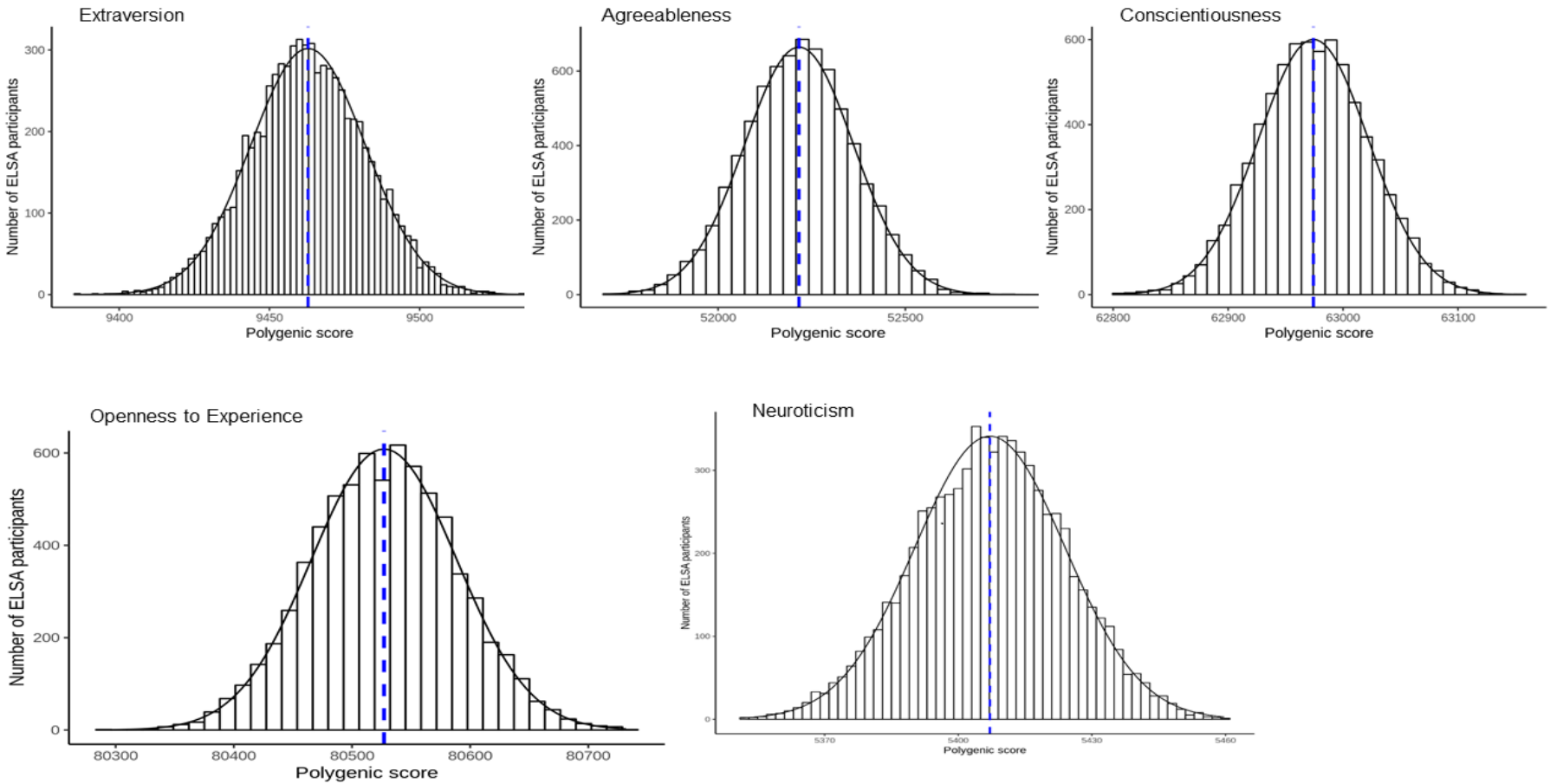
PGS, polygenic score; SE, standard error

Table 3. The sources of the GWAS summary statistics used for these personality types

Phenotype	Consortium	GWAS SNPs	Overlapping with ELSA	GWAS meta-analysis citation	Source of base data
Neuroticism	SSCAG	6,524,432	1,191,041	Okbay et al. (2016)[26]	https://www.thessgac.org/data
Extraversion	GPC	6,941,603	1,218,049	van den Berg et al (2016)[24]	http://www.tweelingenregister.org/GPC/
Agreeableness	GPC	2,305,461	760,918		
Openness	GPC	2,305,738	750,564	de Moor et al. (2012)[25]	http://www.tweelingenregister.org/GPC
Conscientiousness	GPC	2,305,682	750,990		

SSCAG, Social Science Genetic Association Consortium; GPC, Genetics of Personality Consortium

Figure 2. The distributions of the PGSs for Extraversion, Agreeableness, Openness to Experience, Conscientiousness and Neuroticism in the ELSA study



3.3.2. SOCIO-ECONOMIC TRAITS

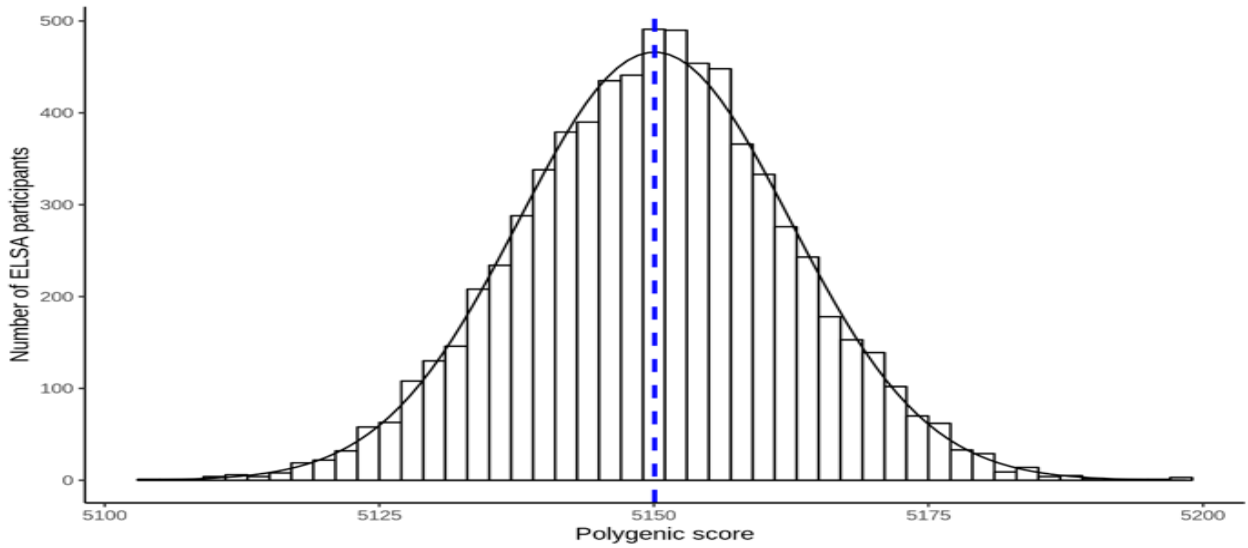
3.3.2.1. Educational Attainment

Educational attainment (EA) is seen as a proxy for educational achievement and to some extend learning [22]. There are two main PGSs for EA available and widely used for research purposes: 1) is based on the GWAS summary statistics developed by Okbay *et al*. (2016), and 2) is based on more recent and much larger GWAS summary statistics provided by Lee *et al*. (2018). To be consistent with the HRS study, in this report these PGSs will be referred to as EA-2 and EA-3, respectively. The detailed methodological information on construction of these PGSs is provided below.

3.3.2.1.1. Educational Attainment - 2 (EA2)

PGSs for EA-2 were created using results from a 2016 study excluding 23andMe results (due to data use agreements)[22]. The meta-analysis included 293,723 individuals in the discovery sample and 111,349 in the replication sample. All samples were restricted to individuals of European descent and whose EA was assessed at or above age 30-year-old. Approximately 9.3 million SNPs were included in the analyses, with the SNPs having been imputed to the 1000 genomes reference panel (1000G)[27]. There were 74 loci that met the genome-wide significance threshold. The educational attainment as measured as years of completed education (i.e., EduYears). This phenotype was constructed by mapping each major educational qualification that can be identified from the survey measure of the cohort to an International Standard Classification of Education (ISCED) category and imputing a years-of-education equivalent for each ISCED category. Study-specific GWASs controlled for the first ten PCs of the genotypic data, a third-order polynomial in age, an indicator for being female, interactions between age and female, and study-specific controls, including dummy variables for major events such as wars or policy changes that may have affected access to education in their specific sample. The distribution of PGS for EA-2 in the ELSA study is depicted in **Figure 3**. The SSGAC GWAS for EA-2 contained 8,146,840 SNPs; of these, 1,316,119 SNPs overlapped with the ELSA target data and were included in the PGSs for EA-2.

Figure 3. Distribution of PGS for EA-2

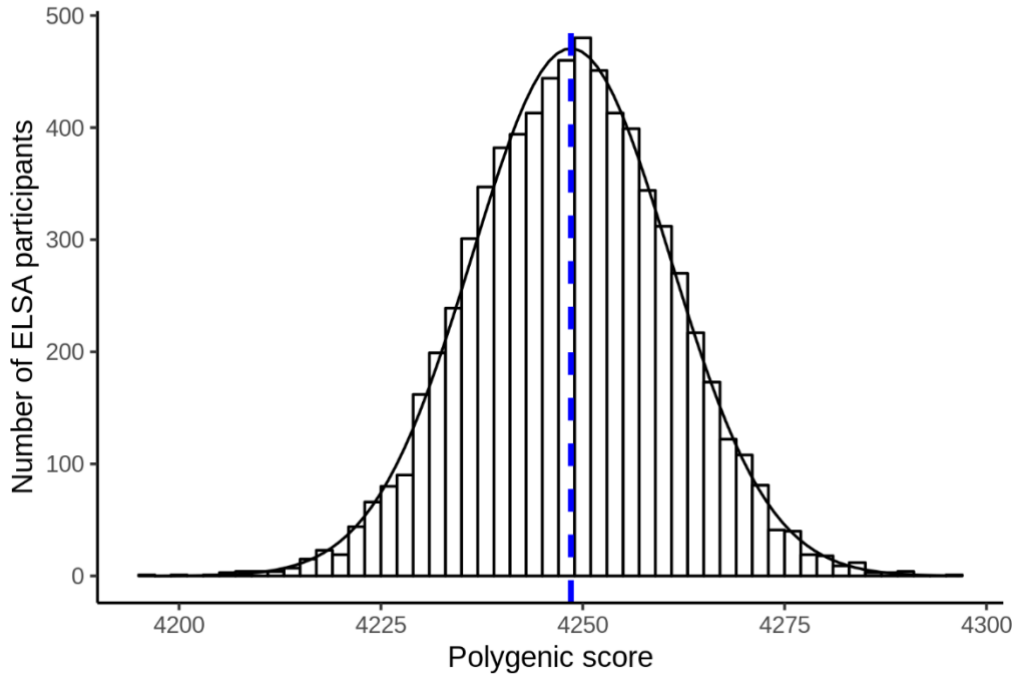


The blue dash line depicts the mean

3.3.2.1.2. Educational Attainment - 3 (EA3)

Similar to EA-2, PGS for EA-3 was created using results from a GWAS carried out by the Social Science Genetic Association Consortium (SSGAC) in 2018 significantly extending the data and number of participates involved [28]. Indeed, the SSGAC GWAS 2018 is an extension of the Okbay's *et al*. (2016) work and was performed on *n*=1125816 individuals across 70 quality-controlled cohorts with all cohorts utilising SNPs imputed to the 1000 genomes reference panel (1000G)[27]. The association analyses in the included datasets were adjusted for sex, birth year, their interaction and 10 PCs of the genetic relatedness matrix. The results showed that a PGS for EA-3 explained around 11% of the variance in educational attainment. The distribution of PGS for EA-3 in the ELSA study is depicted in **Figure 4**. The SSGAC GWAS 2018 contained 10,101,242 SNPs; of these, 1,325,851SNPs overlapped with the ELSA target data and were included in the PGSs for EA.

Figure 4. Distribution of PGS for EA-3



3.3.2.2. Social Deprivation

PGS for social deprivation was created using results from a GWAS carried out using data from UK Biobank [29]. Social deprivation was measured using the Townsend Social Deprivation Index which is a measure of the level of social deprivation in which the participant lives. A total of 112,005 individuals had a Townsend score. The 152,729 blood samples submitted to UK Biobank were genotyped using either the UKBileve array (*n*=49,979) or the UK Biobank axiom array (*n*=102,750). Affymetrix performed genotyping on 33 batches of ~4,700 samples and also conducted the initial quality control procedure on the genotyping data. In addition to the standard quality control procedures applied by the Biobank (http://biobank.ctsu.ox.ac.uk/crystal/refer.cgi?id=155580.), additional quality control was performed for this study. This entailed removing individuals who had non-British ancestry

(within those who self-identified as being British, principal component analysis was used to remove outliers, *n*=32,484), high missingness (n=0), relatedness (*n*=7,948), QC failure in UK Bileve (*n*=187), and gender mismatch (*n*=0). A total of 112,151 individuals remained for further analyses. The UK Biobank interim release was imputed to a reference set which combined the UK10K haplotype and 1000 Genomes Phase 3 reference panels. Association analysis for the social deprivation phenotype was adjusted to control for the effects of age, sex, assessment centre, genotyping batch, genotyping array, and population stratification (using 10 PCs). The PGS for social deprivation contains 1,341,112 SNPs that overlapped between the ELSA genetic database and the GWAS meta-analysis

Figure 5. Distribution of PGS for Social Deprivation

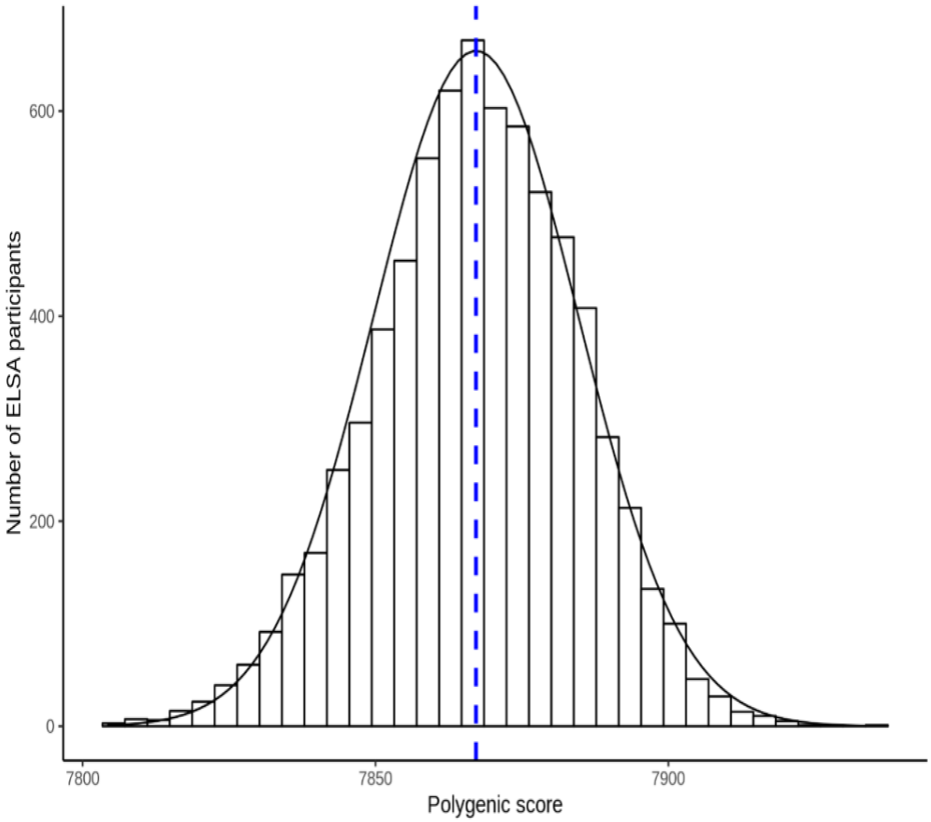


Table 4. The summary statistics for PGSs for each socio-economic trait

PGSs	Sample	Minimum	Maximum	Range	Median	Mean	SE (mean)
Educational Attainment-2	7183	5104.3	5198.5	94.2	5150.2	5150.1	0.15
Educational Attainment-3	7183	4197.0	4296.7	99.7	4248.7	4248.5	0.14
Social Deprivation	7183	7804.2	7934.5	130.3	7867.5	7867.2	0.21

PGS, polygenic score SE, standard error

Table 5. Sources of the GWAS summary statistics used for these socio-economic traits

Phenotype	Consortium	GWAS SNPs	Overlapping with ELSA	GWAS meta-analysis citation	Source of base data
Educational Attainment-2	SSCAG	8,146,840	1,316,119	Okbay et al. (2016)[22]	https://www.thessgac.org/data
Educational Attainment-3	SSCAG	10,101,242	1,325,851	Lee et al. (2018)[28]	On request from the authors
Social Deprivation	-	15,732,391	1,341,112	Hill et al (2016) [29]	https://grasp.nhlbi.nih.gov/FullResults.aspx

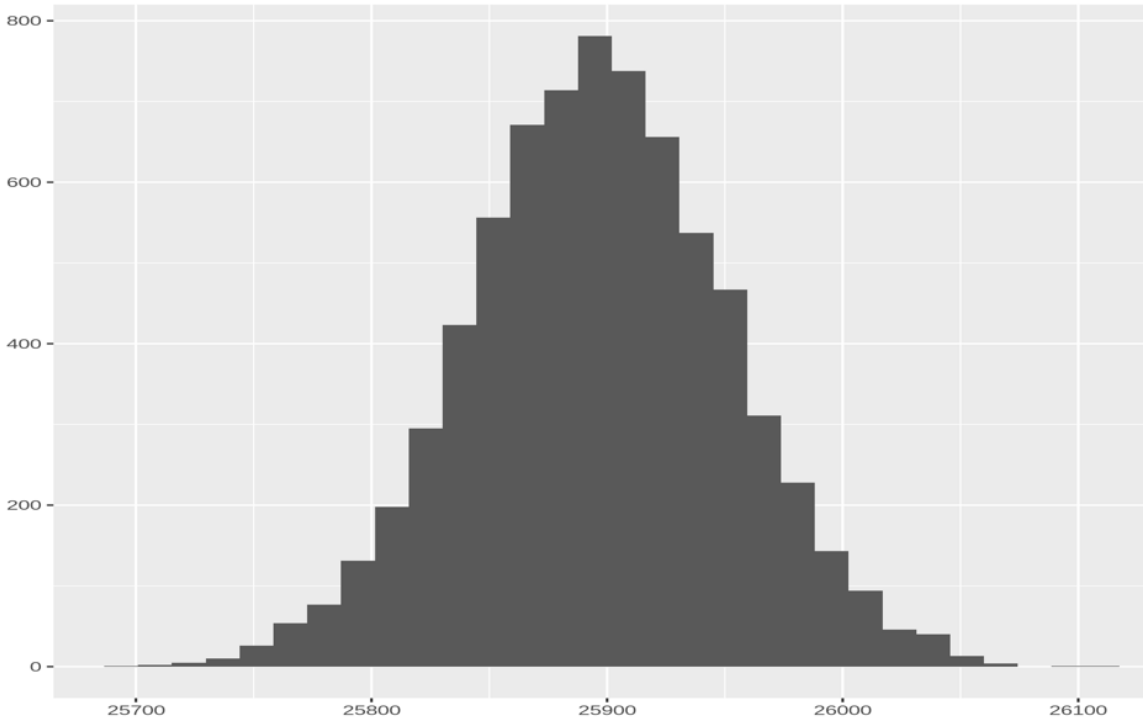
SSCAG, Social Science Genetic Association Consortium

3.3.3. ADULT MENTAL HEALTH AND WELLBEING

3.3.3.1. Alzheimer's disease (2013)

The PGS for Alzheimer's disease (AD) were created using results from a 2013 GWAS conducted by the International Genomics of Alzheimer's Project (IGAP)[30]. The GWAS meta-analysis of AD was conducted across 20 independent studies using data from four international consortia. These included Alzheimer's Disease Genetic Consortium (ADGC), the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium, the European Alzheimer's Disease Initiative (EADI), and the Genetic and Environmental Risk in Alzheimer's Disease (GERAD) Consortium. The stage 1 this meta-analysis included 54,162 participants (n_{cases}=17,008 and $n_{controls}$=37,154) of European decent with a total of 7,055,881 SNPs imputed to 1000 Genomes (2010 release). The stage 2 replication sample included 19,884 participants of European ancestry (n_{cases}=8,572 and $n_{controls}$=11,312) with a total of 11,632 genotyped SNPs. In addition to the *APOE* locus (encoding apolipoprotein E), the two-stage combined discovery and replication GWAS revealed 19 SNPs that reached GWAS significant associations with AD. Adjustment covariates within each contributing cohort included age, sex, and genetic PCs. The distribution of PGS for AD in ELSA is depicted in **Figure 6**. The PGS for AD contains 1,191,420 SNPs that overlapped between the ELSA genetic database and the GWAS meta-analysis.

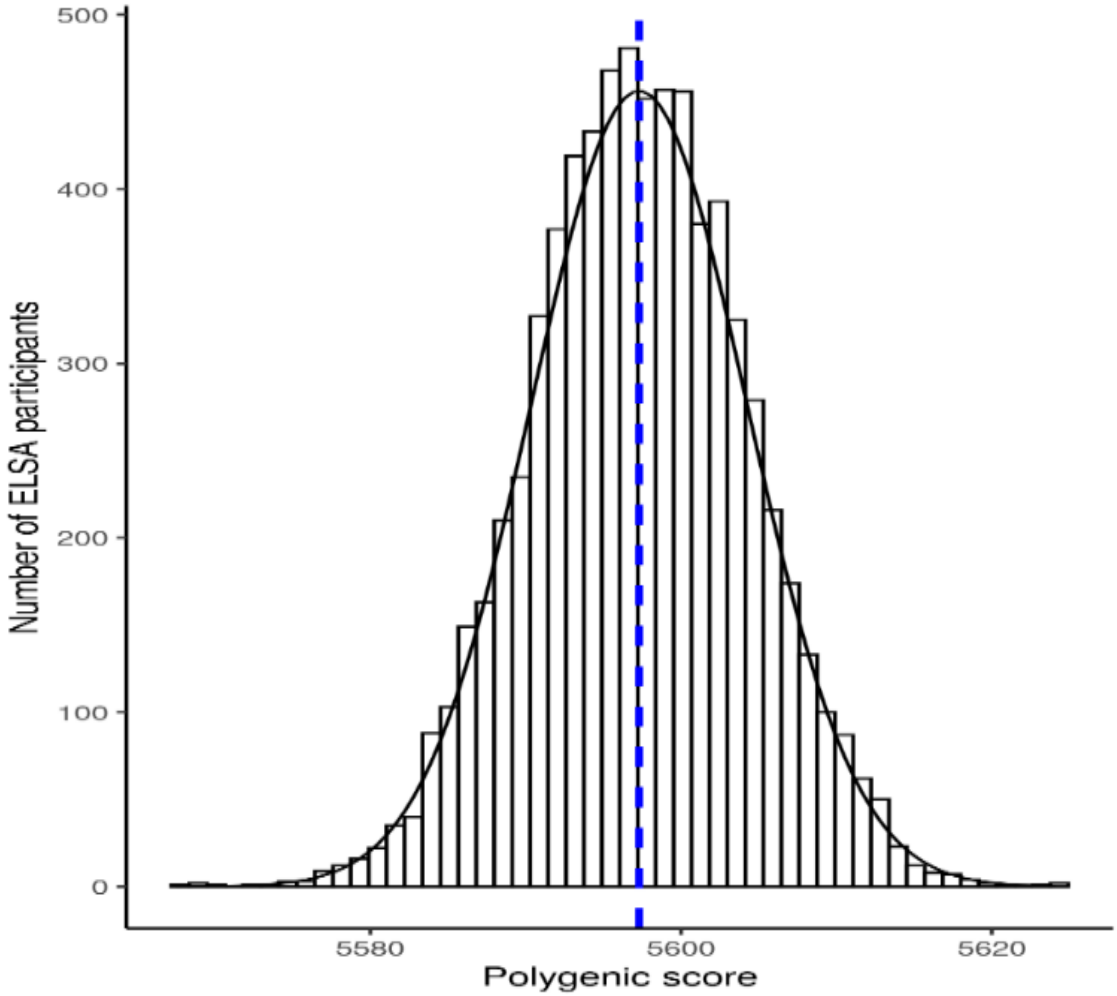
Figure 6. Distribution of PGS for Alzheimer's disease



3.3.3.2. Alzheimer's disease (2019)

The PGS for Alzheimer's disease (AD) (2019) were created using results from a large genome-wide association study of clinically diagnosed AD and AD-by-proxy (71,880 cases, 383,378 controls)[31]. Participants in this study were obtained from multiple sources, including raw data from case-control samples collected by PGC-ALZ and ADSP (made publicly available through dbGaP), summary data from the case-control samples in the IGAP, and raw data from the population-based UKB sample which was used to create a weighted AD-by-proxy phenotype. An additional independent case-control sample (deCODE) was used for replication. AD-by-proxy, based on parental diagnoses, showed strong genetic correlation with AD (r_g = 0.81). Cumulatively, the meta-analysis identified 29 risk loci, implicating 215 potential causative genes. Adjustment covariates within each contributing cohort included age, sex, and genetic PCs. The distribution of PGS for AD (2019) in ELSA is depicted in **Figure 6**. The PGS for AD contains 1712973 SNPs that overlapped between the ELSA genetic database and the GWAS meta-analysis.

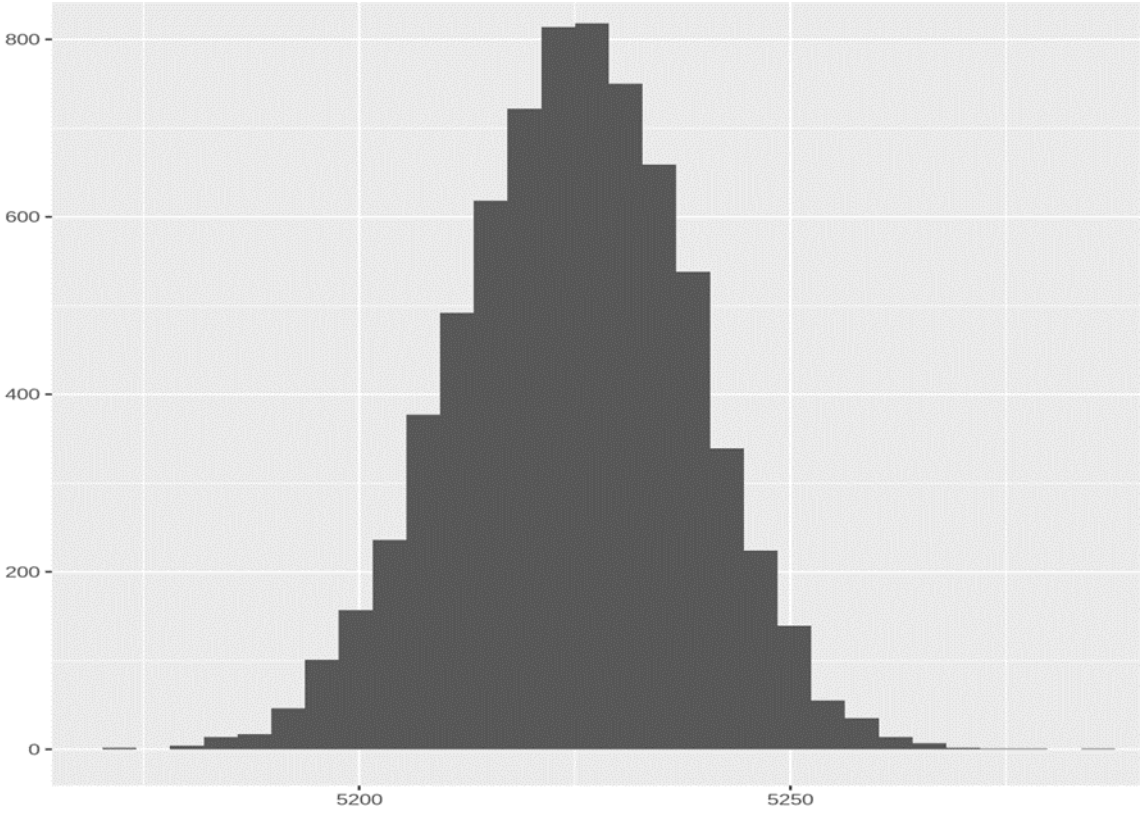
Figure 7. Distribution of PGS for Alzheimer's disease (2019)



3.3.3.3. Depressive Symptoms

PGS for depressive symptoms was created using results from a 2016 GWAS conducted by the Social Science Genetic Association Consortium (SSGAC) as part of their subjective wellbeing GWAS [26]. The SSGAC GWAS included 180,866 individuals and meta-analysed publicly available results from a study performed by the Psychiatric Genomics Consortium (PGC)[32] (n_{cases}=9,240, $n_{controls}$= 9,519) with results from analyses of UK Biobank (UKB) data[33] (n=105,739), and the Resource for Genetic Epidemiology Research on Aging (GERA) Cohort (n_{cases}=7,231, $n_{controls}$=49,316). A replication analysis was also performed using data from 23andMe (n=368,890). To define the phenotype, in UKB, a continuous phenotype measure was used that combined responses to two questions, which asked about the frequency in the past two weeks with which the respondent experienced feelings of unenthusiasm or disinterest and depression or hopelessness. The PGC and GERA cohorts utilised case-control data on major depressive disorder. In the UKB, analyses controlled for the first 15 PCs, indicator variables for genotyping array, sex, indicator variables for age ranges, and sex-by-age interactions [33]. In GERA, analyses controlled for the first four PCs of the genotypic data, sex, and 14 indicator variables for age ranges. The PGC included controls for five PCs, sex, age, and cohort fixed effects [32]. The distribution of PGS for Depressive Symptoms in ELSA is depicted in **Figure 8**. GWAS summary statistics contained 6524474 SNPs; of these, 1187563 SNPs overlapped with the ELSA genetic database and were included in the PGS for depressive symptoms phenotype.

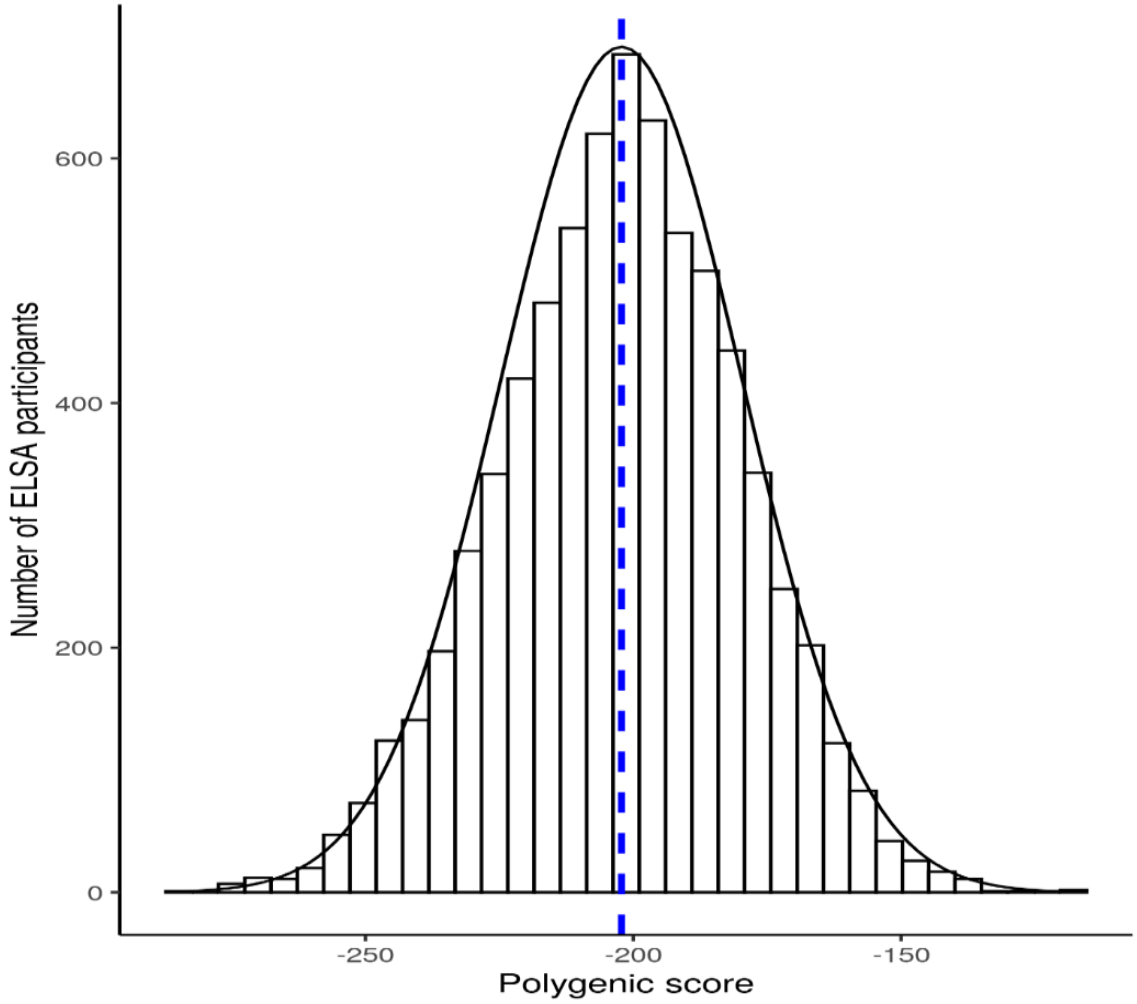
Figure 8. Distribution of PGS for Depressive Symptoms



3.3.3.4. Major Depressive Disorder (2018)

The PGSs for Major Depressive Disorder (MDD) were created using results from a 2018 GWAS conducted by the MDD working group of the Psychiatric GWAS Consortium (PGC)[34]. At the time we were preparing the PGS for MDD, the GWAS meta-analysis files are available on the PGC website: http://www.med.unc.edu/pgc/results-and-downloads. PGC conducted a genome-wide association meta-analysis based in 135,458 cases and 344,901 controls which identified 44 independent and significant loci. MDD cases were required to have a DSM-IV lifetime MDD diagnosis that was collected by a clinician using structured interviews or clinician-administered DSM-IV checklists. Most controls were randomly selected and screened for lifetime MDD. The distribution of PGS for MDD in ELSA is depicted in **Figure 9**. GWAS summary statistics contained 8,483,301 SNPs; of these, 1197733 SNPs overlapped with the ELSA genetic database and were included in the PGS for MDD.

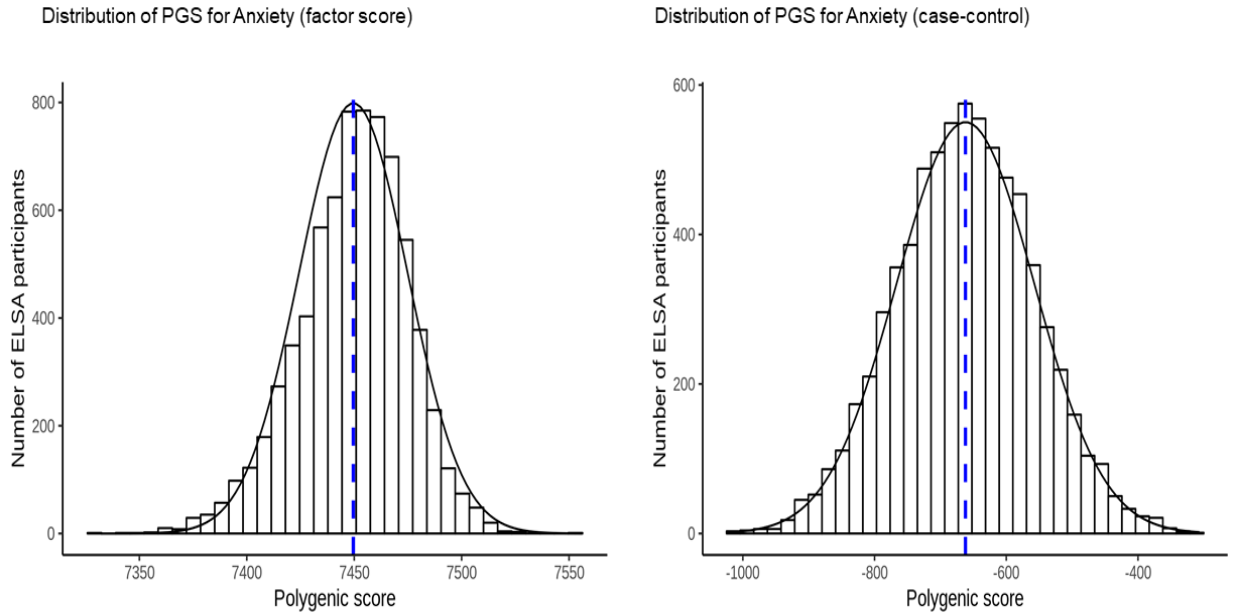
Figure 9. Distribution of PGS for MDD 2018



3.3.3.5. Anxiety (case-control, factor score)

Anxiety Disorders (AD) included generalized AD, panic disorder and phobias. The PGs for the GAD was calculated using the GWAS meta-analysis which combined results across the nine studies participating in the Anxiety NeuroGenetics STudy (ANGST) Consortium for over 18000 unrelated individuals[35]. The combined case-control meta-analysis included *n*=17,310 and the continuous factor score GWAS included *n*=18,186. All cohorts imputed SNPs to the 1000 Genomes Project references data (release v3, March 2012) and approximately 6.5 million SNPs were included in the combined meta-analysis. The regression analyses were adjusted for sex and age at interview, as they were significant predictors of the phenotypes. Ancestry principal components were estimated for each sample and included on a sample-by-sample basis depending on their correlation with the phenotypes. The authors conducted two types of analyses in each sample based on complementary approaches to modelling the comorbidity and common genetic risk across the ADs: (1) CC comparisons, in which cases were designated as having 'any AD' versus supernormal controls, and (2) quantitative FS estimated for every subject in the sample using confirmatory factor analysis. The distribution of PGS for Anxiety (case-control, factor score) in ELSA is depicted in **Figure 10**. From the ANGST meta-analysis, 1,137,311 SNPs overlapped with the ELSA genetic database and were included in the PGS for Anxiety (factor score) phenotype and 1,068,194 SNPs overlapped with the ELSA genetic database and were included in the PGS for Anxiety (case-control) phenotype.

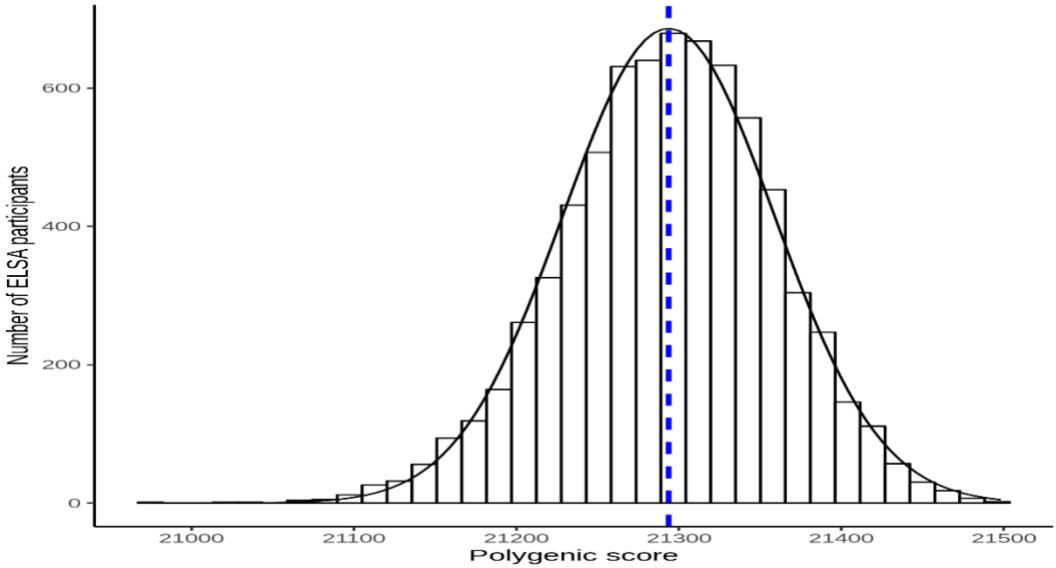
Figure 10. Distribution of PGS for Anxiety (case-control, factor score)



3.3.3.6. Schizophrenia (2014)

The PGSs for schizophrenia (2014) were created using results from a 2014 GWAS conducted by the Schizophrenia Working Group of the Psychiatric Genomics Consortium (PGC)[36]. The schizophrenia GWAS combined meta-analysis included 36,989 cases and 113,075 controls (N=152,805) and identified 128 independent associations spanning 108 conservatively defined loci that meet genome-wide significance, 83 of which have not been previously reported. The replication sample consisted of 1,513 cases and 66,236 controls. After quality control, around 9.5 million SNPs were included in the analyses. Genetic principal components and study identifiers were included as covariates. The distribution of PGS for Schizophrenia in ELSA is depicted in **Figure 11**. The PGS contain 1,278,742 SNPs that overlapped between the ELSA genetic database and the GWAS meta-analysis; these SNPs were included in the PGS for Schizophrenia (2014).

Figure 11. Distribution of PGS for Schizophrenia (2014)



3.3.3.7. Schizophrenia (2020)

The PGSs for schizophrenia (2020) were created using results from a 2020 GWAS conducted by the Schizophrenia Working Group of the Psychiatric Genomics Consortium (PGC). The schizophrenia GWAS combined meta-analysis included 69,369 people with schizophrenia and 236,642 controls and identified 270 independent associations spanning 130 genes. The PGS contain 1,862,381 SNPs that overlapped between the ELSA genetic database and the GWAS meta-analysis; these SNPs were included in the PGS for Schizophrenia (2020).

3.3.3.8. Bipolar disorders (2019)

GWAS sample comprises 32 cohorts from 14 countries in Europe, North America and Australia, totalling 20,352 cases and 31,358 controls of European descent[37]. Cases were required to meet international consensus criteria (DSM-IV or ICD-10) for a lifetime diagnosis of BD established using structured diagnostic instruments from assessments by trained interviewers, clinician-administered checklists, or medical record review. In most cohorts,

controls were screened for the absence of lifetime psychiatric disorders and randomly selected from the population. Variant dosages were imputed using the 1000 Genomes reference panel, retaining association results for 9,372,253 autosomal variants with imputation quality score INFO > 0.3 and minor allele frequency (MAF) ≥1% in both cases and controls. Logistic regression of case status on imputed variant dosage was performed using genetic ancestry covariates. The resulting genomic inflation factor (λ_{GC}) was 1.23, 1.01 when scaled to 1,000 cases and 1,000 controls (λ_{1000}). The linkage disequilibrium (LD) score regression intercept was 1.021 (s.e.m. = 0.010), and the attenuation ratio of 0.053 (s.e.m. = 0.027) was non-significant, indicating that the observed genomic inflation is indicative of polygenicity rather than stratification or cryptic population structure. The LD score regression SNP heritability estimates for BD were 0.17–0.23 on the liability scale assuming population prevalence of 0.5–2%.

3.3.3.9. Bipolar disorders (2021)

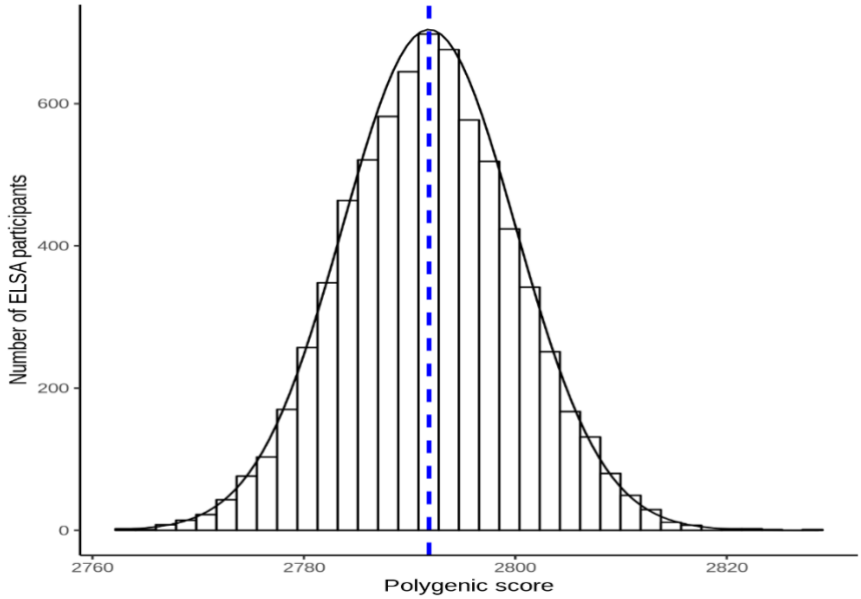
The GWAS meta-analysis sample comprised 57 cohorts collected in Europe, North America and Australia, totalling 41917 BD cases and 371549 controls of European descent[38]. The total effective n, equivalent to an equal number of cases and controls in each cohort ($4 \times n_{cases} \times n_{controls}/(n_{cases} + n_{controls})$), is 101962. Cases were required to meet international consensus criteria (DSM-IV, ICD-9, or ICD-10) for a lifetime diagnosis of BD, established using structured diagnostic instruments from assessments by trained interviewers, clinician-administered checklists, or medical record review. In most cohorts, controls were screened for the absence of lifetime psychiatric disorders and randomly selected from the population. he GWAS meta-analysis identified 64 independent loci associated with BD at genome-wide significance ($P < 5 \times 10^{-8}$). Using linkage disequilibrium score regression (LDSC), the h^2_{SNP} of BD was estimated to be 18.6% (s.e. = 0.008, $P = 5.1 \times 10^{-132}$) on the liability scale, assuming a BD population prevalence of 2%, and 15.6% (s.e. = 0.006, $P = 5.0 \times 10^{-132}$) assuming a population prevalence of 1%. The genomic inflation factor (λ_{GC}) was 1.38 and the LDSC intercept was 1.04 (s.e. = 0.01, $P = 2.5 \times 10^{-4}$). While the intercept has frequently been used as an indicator of confounding from population stratification, it can rise above 1 with increased sample size and heritability. The attenuation ratio - (LDSC intercept − 1)/(mean of association chi-square statistics − 1) - which is not subject to these limitations, was 0.06 (s.e. = 0.02), indicating that the majority of inflation of the GWAS test statistics was due to polygenicity. Of the 64 genome-wide significant loci, 33 are novel discoveries (that is, loci not overlapping with any locus previously reported as genome-wide significant for BD). Novel loci include the major histocompatibility complex (MHC) and loci previously reaching genome-wide significance for other psychiatric disorders, including 10 for schizophrenia, 4 for major depression and three for childhood-onset psychiatric disorders or problematic alcohol use.

3.3.3.10. Subjective Well-Being

PGSs for subjective wellbeing (SWB) were created using results from a 2016 GWAS conducted by the Social Science Genetic Association Consortium (SSGAC)[26]. The subjective wellbeing analyses included 298420 European ancestry individuals in the discovery sample. Genome-wide significant SNPs were identified in 3 loci. The phenotype measure was life satisfaction, positive affect, or in some cohorts a measure combining both. Approximately 9.3 million SNPs were included in the analyses, with cohorts utilising SNPs imputed to the 1000 genomes reference panel (1000G) or the HapMap 2 Panel. Adjustments for age, age^2,

sex, and four PCs from the genotypic data were included in study specific GWAS association analyses. Cohorts were also asked to include any study-specific covariates such as study site or batch effects. The distribution of PGS for SWB in ELSA is depicted in **Figure 12**. GWAS summary statistics contained 2268674 SNPs; of these, 748500 SNPs overlapped with the ELSA genetic database and were included in the PGS for SWB phenotype.

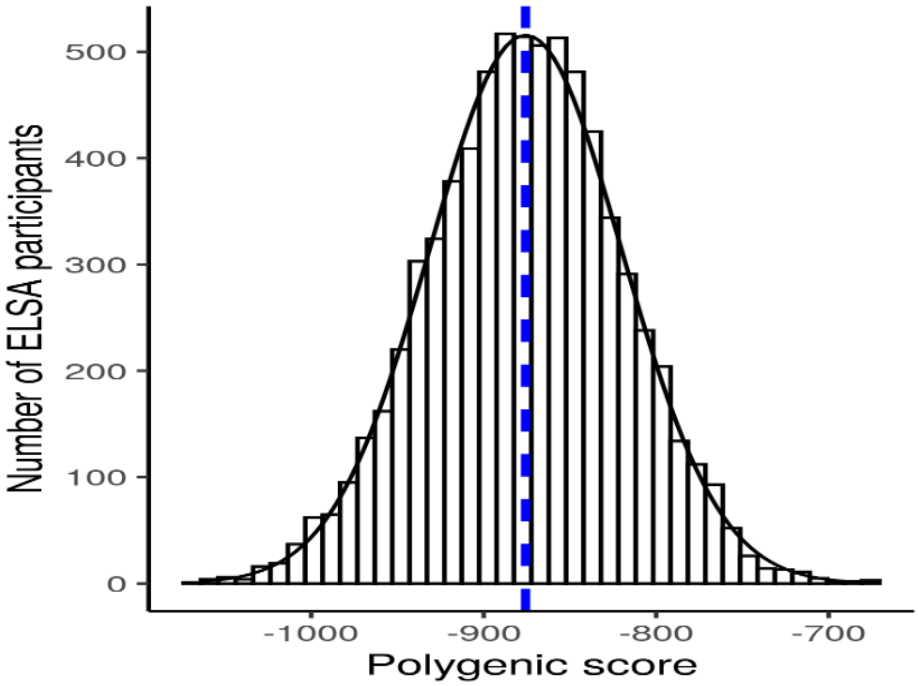
Figure 12. Distribution of PGS for Subjective Well-Being



3.3.3.11. Attention deficit hyperactivity disorder (2019)

PGSs for ADHD were calculated using a genome-wide association meta-analysis of 20,183 ADHD cases and 35,191 controls were collected from 12 cohorts that identifies variants surpassing genome-wide significance in 12 independent loci[39]. These samples included a population-based cohort of 14,584 cases and 22,492 controls from Denmark collected by the Lundbeck Foundation Initiative for Integrative Psychiatric Research (iPSYCH) and 11 European, North American, and Chinese cohorts aggregated by the Psychiatric Genomics Consortium (PGC). ADHD cases in iPSYCH were identified from the national Psychiatric Central Research Register psychiatric and diagnosed by psychiatrists at a psychiatric hospital according to ICD10 (F90.0) and genotyped using Illumina PsychChip. In total, 304 genetic variants in 12 loci surpassed the threshold for genome-wide significance. GWAS summary statistics contained 8094094 SNPs; of these, 1099981 SNPs overlapped with the ELSA genetic database and were included in the PGS for ADHD phenotype.

Figure 13. Distribution of PGS for ADHD (2019)



3.3.3.12. Autism spectrum disorder (2016)

PGSs for Autism spectrum disorder (ASD) were calculated using meta-analyses that combined data from 14 independent cohorts contributed by eight academic studies; each contributing site confirmed all affected individuals had an ASD diagnosis. Where data permitted, individuals assessed at under 36 months of age or if there was any evidence of diagnostic criteria not being met from either the Autism Diagnostic Interview-Revised (ADI-R) or the Autism Diagnostic Observation Schedule (ADOS), were excluded. The primary meta-analysis (Worldwide ancestry (WW)) was based on data from 7387 ASD cases and 8567 controls. An independent replication of the results using summary GWAS findings from two additional sources (i.e., the Danish iPSYCH Project (7783 ASD cases and 11359 controls) and a combined deCODE Collection (from Iceland plus a collection of individuals from Ukraine, Georgia and Serbia) and the 'Study to Explore Early Development' (SEED) (1369 ASD cases and 137308 controls)) was carried out. GWAS summary statistics contained 6440258 SNPs; of these, 1094347 SNPs overlapped with the ELSA genetic database and were included in the PGS for ASD phenotype.

3.3.3.13. Loneliness (2016)

PGCs for loneliness were created using results from a 2016 GWAS conducted by the Psychiatric Genomics Consortium utilising genotypic and phenotypic data from 10 760 individuals aged ⩾50 years that were collected by the Health and Retirement Study (HRS) to perform the first genome-wide association study of loneliness[40]. No associations reached genome-wide significance (P_{GWAS} >5×10^{-8}). Furthermore, none of the previously published associations between variants within candidate genes (BDNF, OXTR, RORA, GRM8, CHRNA4, IL-1A, CRHR1, MTHFR, DRD2, APOE) and loneliness were replicated

(P_{GWAS}>0.05), despite our much larger sample size. Cohorts were also asked to include any study-specific covariates such as study site or batch effects. GWAS summary statistics contained 5768558 SNPs; of these, 1055906 SNPs overlapped with the ELSA genetic database and were included in the PGS for Loneliness (2016) phenotype.

3.3.3.14. Loneliness (2018)

PGSs for Loneliness (2018) were calculated using summary which were based on data from the May 2017 release of imputed genetic data from UK Biobank[41] which included 40 M imputed variants from the HRC reference panel. Individuals clustered into this group who self-identified by questionnaire as being of an ancestry other than white European were excluded. After application of QC criteria, a maximum of 452,302 individuals were available for analysis with genotype and phenotype data. Loneliness was derived from self-reported answers to questions directed via assessment centre touchscreen. The data from three related questions was assessed loneliness and social isolation–(1) 'Do you often feel lonely?', to which individuals answered 'yes' (recorded as cases) or 'no' (controls), (2) A composite variable based on the questions 'Including yourself, how many people are living together in your household?' and "How often do you visit friends or family or have them visit you?' (cases were defined as those who lived alone and who indicated that they either never visited or had no friends or family outside their household; controls were defined as those who either did not live alone, or had friends who visited at least once a week) and (3) A variable representing quality of social interactions 'How often are you able to confide in someone close to you?' (cases were defined as those who answered 'Never or almost never', controls were defined as those who answered 'Almost daily'). GWAS summary statistics contained 7,745,443 SNPs; of these, 1342866 SNPs overlapped with the ELSA genetic database and were included in the PGS for Loneliness (2018) phenotype.

Table 6. Presents the summary statistics for PGS for each adult mental health and wellbeing outcome

PGS	Sample Size	Minimum	Maximum	Range	Median	Mean	SE (mean)
Alzheimer's disease (2013)	7183	25696	26112.5	416.5	25896	25896.2	0.64
Alzheimer's disease (2019)	7183	5567.5	5624.1	56.6	5597.3	5597.3	0.08
Anxiety (case-control)	7183	-1010.4	-306.6	703.9	-661.2	-662.0	1.23
Anxiety (factor score)	7183	7332.1	7556.0	223.9	7451.6	7449.6	0.30
Attention deficit hyperactivity disorder (2019)	7183	-1064.0	-670.7	393.3	-875.2	-875.7	0.66
Autism spectrum disorder (2016)	7183	30140.6	30808.6	668.0	30429.3	30427.9	1.02
Bipolar disorders (2019)	7183	-1611.5	-1153.6	457.8	-1377.0	-1377.1	0.69
Bipolar disorders (2021)	7183	96558.4	98588.1	2029.7	97642.6	97634.8	3.20
Depressive Symptoms (DS)	7183	5170.3	5283.8	113.5	5225.0	5224.7	0.16
Loneliness (2016)	7183	34514.4	35006.3	491.9	34736.9	34737.4	0.81
Loneliness (2018)	7183	1579.1	1607.3	28.3	1594.6	1594.5	0.05
Major Depressive Disorder (2018)	7183	-283.8	-116.9	166.9	-201.8	-202.2	0.26
Schizophrenia (2014)	7183	20976.9	21498.2	521.3	21295.7	21293.7	0.77
Schizophrenia (2020)	7183	-3248.9	-2628.6	620.3	-2811.5	-2829.7	1.06
Subjective Well-Being	7183	2763.7	2828.7	6465.0	2791.8	2791.8	0.10

PGS, polygenic score; SE, standard error

Table 7. Sources of the GWAS summary statistics used for PGS for each adult mental health and wellbeing outcome

Phenotype	Consortium	GWAS SNPs	Overlapping with ELSA	GWAS meta-analysis citation	Source of base data
ADHD	PGC	8,094,094	1,099,981	Brainstorm Consortium et al (2019)[39]	https://www.med.unc.edu/pgc/results-and-downloads
Alzheimer's disease (2013)	IGAP	7,055,881	1,191,420	Lambert et al. (2013)[30]	http://web.pasteur-lille.fr/en/recherche/u744/igap/igap_download.php
Alzheimer's disease (2019)	-		1,712,973	Jansen et al (2019)[31]	https://ctg.cncr.nl/software/summary_statistics
Anxiety Disorders (case-control)	ANGST		1,068,194	Otowa et al. (2016)[35]	https://www.med.unc.edu/pgc/results-and-downloads
Anxiety Disorders (factor score)	ANGST	6,306,612	1,137,311	Otowa et al. (2016)[35]	https://www.med.unc.edu/pgc/results-and-downloads
Autism Spectrum Disorders	PGC	6,440,258	1,094,347	Psychiatric Genomics Consortium (2017)[42]	https://www.med.unc.edu/pgc/results-and-downloads
Bipolar disorders (2019)	PGC	12,369,815	1,228,480	Stahl et al. (2019)[37]	https://www.med.unc.edu/pgc/results-and-downloads
Bipolar disorders (2021)	PGC	7,608,183	5,901,703	Mullins et al. (2021)[38]	https://www.med.unc.edu/pgc/results-and-downloads
Depressive symptoms	SSCAG	6,524,474	1,187,563	Okbay e al. (2016)[26]	https://www.thessgac.org/data
Insomnia Complaints	-	12,444,915	803,361	Hammerschlag et al (2017)[43]	http://ctg.cncr.nl/software/summary_statistics
Loneliness (2017)	PGC	5,768,558	1,055,906	Gao et al (2017)[40]	https://www.med.unc.edu/pgc/results-and-downloads
Loneliness (2018)	PGC	7,745,443	1,342,866	Day et al (2018)[41]	https://www.med.unc.edu/pgc/results-and-downloads
Major Depressive Disorder (2018)	PGC	8,483,301	1,197,733	Wray et al (2018)[34]	https://www.med.unc.edu/pgc/results-and-downloads
Schizophrenia (2014)	PGC	9,444,230	1,278,742	Ripke et al. (2014)[36]	https://www.med.unc.edu/pgc/results-and-downloads (scz2.snp.results.txt.gz)
Schizophrenia (2020)	PGC	7,564,369	1,862,381	Psychiatric Genomics Consortium et al (2020)	https://www.med.unc.edu/pgc/results-and-downloads
Subjective Well-Being	SSCAG	2,268,674	748,500	Okbay et al (2016)[26]	On request from the authors

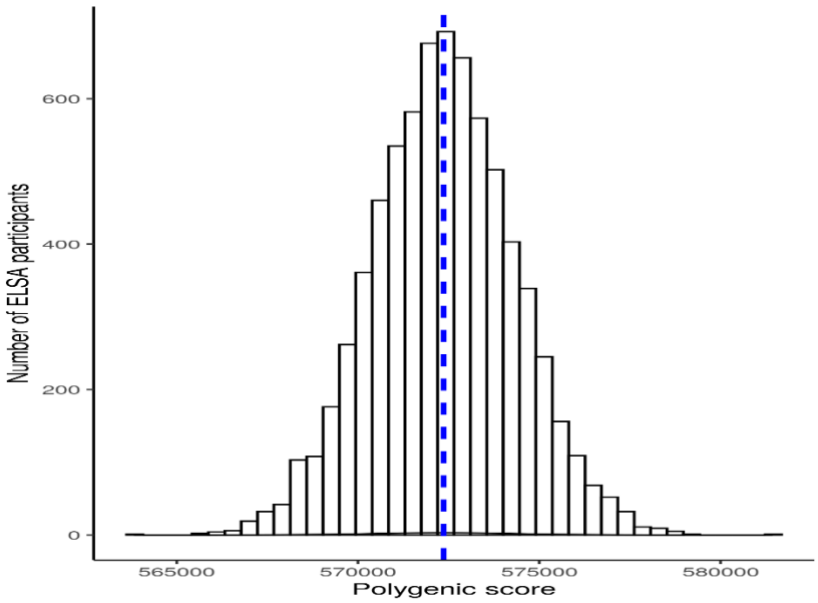
PGC, Psychiatric Genomics Consortium; IGAP, International Genomics of Alzheimer's Project; ANGST, Anxiety NeuroGenetics STudyConsortium; SSCAG, Social Science Genetic Association Consortium;

3.3.4. CHILDHOOD EXPERIENCES

3.3.4.1. Aggressive behaviour in childhood (2015)

PGCs for Aggressive behaviour in childhood was calculated using the summary statistics from GWAS performed in the framework of the Early Genetics and Life course Epidemiology (EAGLE) consortium (http://research.lunenfeld.ca/eagle/)[44]. Nine cohorts contributed data for the total meta-analysis (N = 18,988 children, mean age = 8.44 years, SD = 4.16). In all cohorts that were included in this meta-analysis, well-validated questionnaires assessing aggressive behaviour in children were mailed to parents of children. In eight out of the nine cohorts, maternal ratings of children's aggressive behaviour were obtained. In GINI + LISA, the majority (>80%) of the questionnaires were filled in by the mother. Aggressive behaviour was measured on a continuous scale (with higher scores indicating more aggressive behaviour). The distribution of PGS for Aggressive behaviour in childhood in ELSA is depicted in **Figure 14**. GWAS summary statistics contained 2,188,528 SNPs; of these, 722,659 SNPs overlapped with the ELSA genetic database and were included in the PGS for Aggressive behaviour in childhood phenotype.

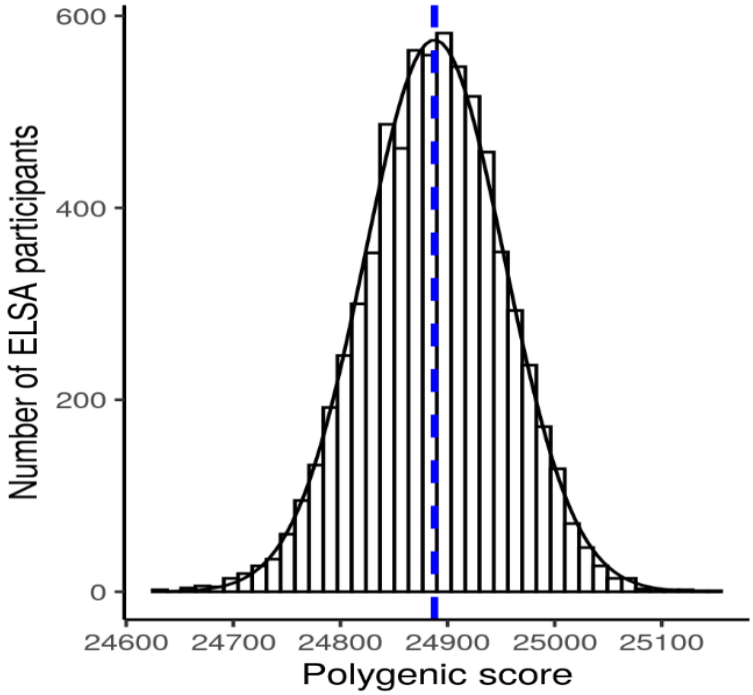
Figure 14. Distribution of PGS for Aggressive behaviour in childhood (2015)



3.3.4.2. Pre-school internalising (2014)

PGSs for internalizing problems was calculated using the summary statistics from GWAS that collated data from 3 cohorts (total N=4,596 children) in which Preschool internalizing problems was assessed with the same instrument, the Child Behavior Checklist[45]. The results showed that Genome-wide SNPs explained 13% to 43% of the total variance in preschool internalizing problems. The meta-analysis did not yield a genome-wide significant signal but was suggestive for the PCSK2 gene located on chromosome 20p12.1. The distribution of PGS for Preschool internalizing problems in ELSA is depicted in **Figure 15**. GWAS summary statistics contained 2,268,674 SNPs; of these, 798,070 SNPs overlapped with the ELSA genetic database and were included in the PGS for Preschool internalizing problems phenotype.

Figure 15. Distribution of PGS for Pre-school internalising (2014)



3.3.4.3. Childhood trauma

PGSs for Childhood trauma were calculated using summary statistics from a meta-analysis that combined 19 GWASs comprising subjects of European ancestry only. The discovery dataset consisted of 124,711 individuals with available childhood maltreatment data from the UK Biobank (UKBB) and the replication sample comprised 26,290 individuals - a subset of the PGC-PTSD Freeze 1.5 dataset (PGC1.5). For the childhood maltreatment phenotype, Childhood Trauma Questionnaire (CTQ) scores on physical, sexual, and emotional abuse were obtained from the participating studies. From this, an overall childhood maltreatment count score of 0–3 was constructed, based on a count of the three abuse categories listed above. An individual was considered to have endorsed a childhood abuse category if they scored in the moderate to extreme range for that particular category, per established cut-offs. If CTQ data were not available, the event assessment during childhood (occurring before 18 years of age) that was most validated for that particular study was obtained, providing a count

of the total number of different categories of reported childhood events (e.g. physical, sexual, or severe emotional abuse) along with the range of possible scores for the measure. The distribution of PGS for Childhood trauma in ELSA is depicted in **Figure 16**. A total of 1143861 SNPs overlapped with the ELSA genetic database with the GWAS summary statistics and were included in the PGS for Childhood trauma phenotype.

Figure 16. Distribution of PGSs for childhood trauma

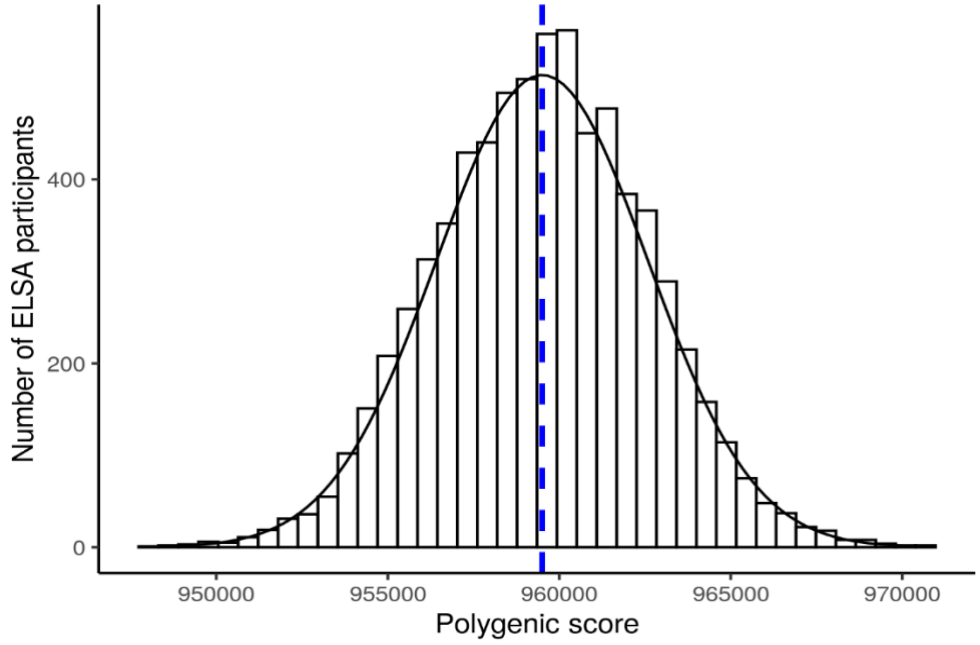


Table 8. The summary statistics for PGS for childhood experiences

PGS	Sample Size	Minimum	Maximum	Range	Median	Mean	SE (mean)
Childhood trauma	7183	948116	970771	22655	959548.0	959501.8	36.3
Pre-school internalising (2014)	6991	24636.3	25153.6	1236.4	24887.0	24866.9	1.66
Aggressive behaviour in childhood (2015)	7183	563819	581439	17620	572363.0	572363.0	23.0

PGS, polygenic score; SE, standard error

Table 9. Sources of the GWAS summary statistics used for PGS childhood experiences

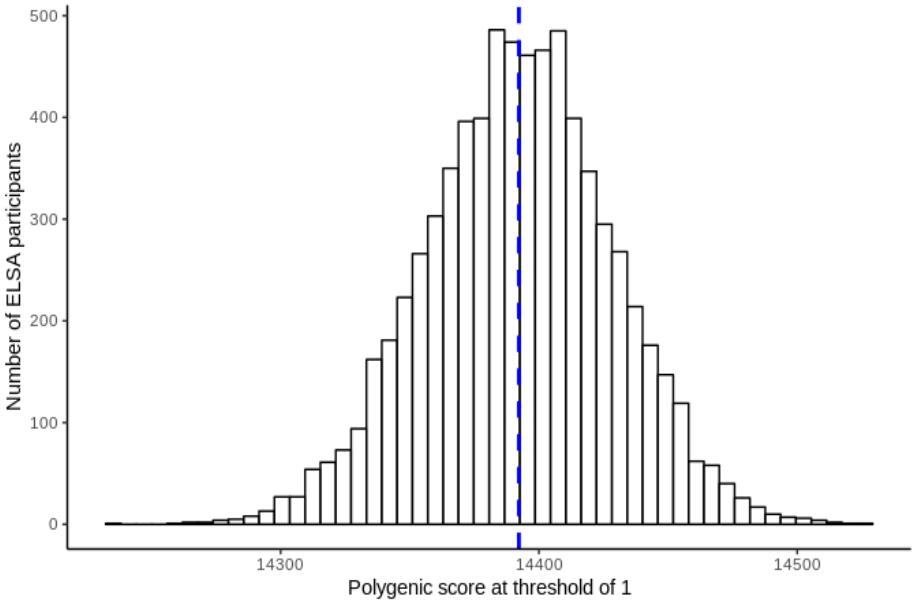
Phenotype	Consortium	GWAS SNPs	Overlapping with ELSA	GWAS meta-analysis citation	Source of base data
Aggressive behaviour in childhood	EAGLE	2,188,528	722,659	Pappa et al (2015)[44]	https://www.wikigenes.org/e/art/e/348.html
Preschool Internalizing Problems	EAGLE	2,821,734	798,070	Benke et al (2014)[45]	https://www.wikigenes.org/e/art/e/348.html
Childhood trauma	PGC	8,031,871	1,143,861	Dalvie et al (2020)[46]	https://www.med.unc.edu/pgc/results-and-downloads

3.3.5. PHYSICAL HEALTH OUTCOMES

3.3.5.1. Coronary Artery Disease (2011)

PGS for coronary artery disease (CAD) was created using results from a 2011 study conducted by the Coronary ARtery DIsease Genome wide Replication and Meta-analysis (CARDIoGRAM) consortium[47]. The GWAS meta-analysis consisted of 14 studies with a total of 22,233 individuals with CAD (cases) and 64,762 without CAD (controls) of European descent imputed to the HapMap3 CEU panel. Replication was performed in a sample of 56,682 individuals (approximately half cases and half controls). This analysis identified 13 loci newly associated with CAD at $P_{GWAS} < 5 \times 10^{-8}$ which had risk allele frequencies ranging from 0.13 to 0.91 and were associated with a 6% to 17% increase in the risk of CAD per allele. The results of these analyses also confirmed the association of 10 of 12 previously reported CAD loci. Study-specific GWAS adjusted for age of onset (cases) or age of recruitment (controls), gender, and genetic PCs. The distribution of PGS for CAD in ELSA is depicted in **Figure 17**. The PGS contain 783,413 SNPs that overlapped between the ELSA genetic database and the GWAS meta-analysis; these SNPs were included in the PGS for CAD.

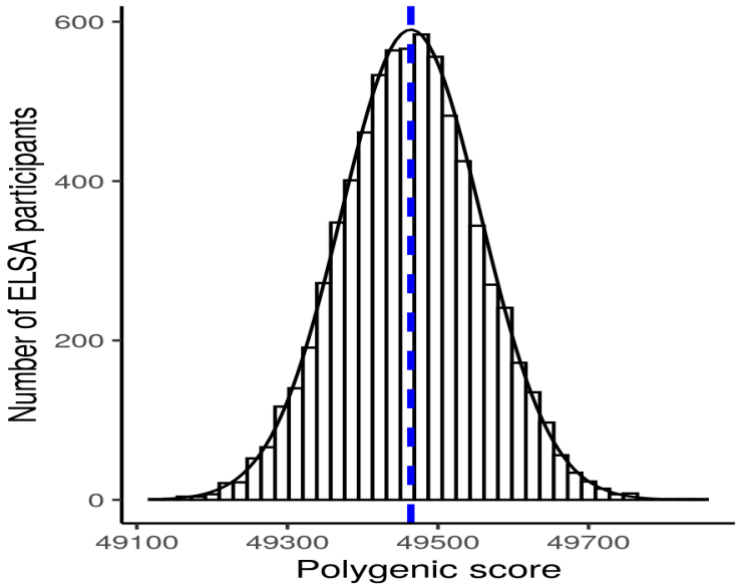
Figure 17. Distribution of PGS for Coronary Artery Disease (2011)



3.3.5.2. Coronary Artery Disease (2018)

PGSs for Coronary Artery Disease (CAD) were calculated using the summary statistics from the largest genome-wide association study of CAD among individuals with diabetes to date[48]. Here, the primary data set contained 15,666 individuals with diabetes of white British ancestry (3,968 CAD cases and 11,698 controls). The outcome of interest was CAD based on UK Biobank's baseline assessment verbal health interview, combined with linked data from hospital admissions and death registries. In total, 9,087,334 autosomal variants with a quality score (IMPUTE2 information metric) >0.8 and a minor allele count ≥30 in cases and controls were analysed in a series of logistic regression models adjusting for age, sex, the first 20 principal components and genotyping batch (3 levels; the UK BiLEVE, UK Biobank release 1 and 2). The association tests were performed in PLINK v2.00 (www.cog-genomics.org/plink/2.0/)) using dosages (additive coding). Overall, this study provided evidence that the genetic mechanisms underlying CAD in participants with diabetes are similar to those in individuals without diabetes. The distribution of PGS for CAD (2018) in ELSA is depicted in **Figure 18**. The PGS contain 1,349,976 SNPs that overlapped between the ELSA genetic database and the GWAS meta-analysis; these SNPs were included in the PGS for CAD (2018).

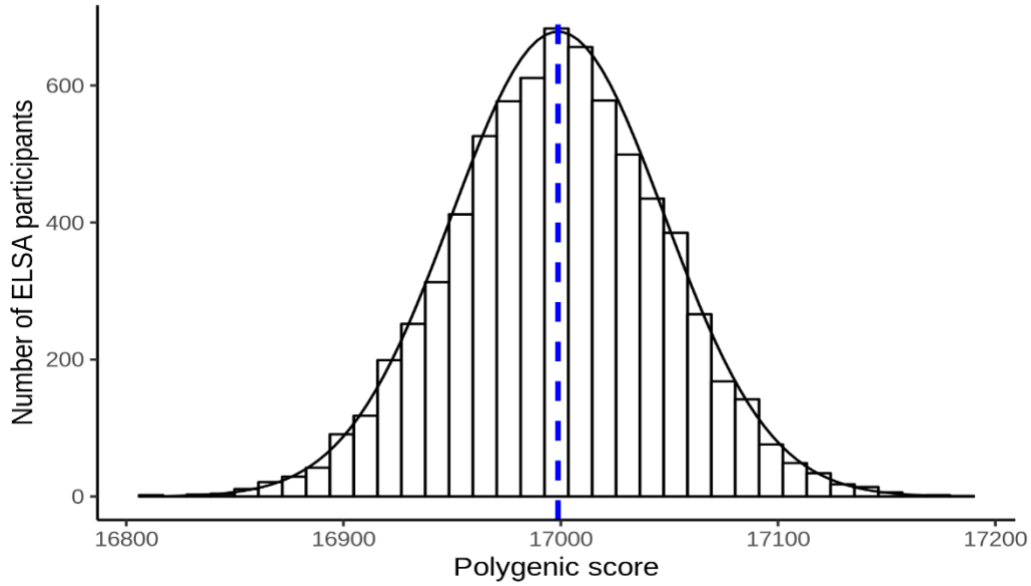
Figure 18. Distribution of PGS for Coronary Artery Disease (2018)



3.3.5.3. Type II Diabetes (2012)

PGSs for Type II Diabetes (T2D) were created using GWAS meta-analysis results from a 2012 study conducted by the DIAbetes Genetics Replication and Meta-analysis (DIAGRAM) Consortium[49]. The stage one (discovery) meta-analysis consists of 12,171 T2D cases and 56,862 controls across 12 GWAS from European descent populations. The stage two (replication) meta-analysis consisted of 22669 cases and 58119 controls, including 1178 cases and 2472 controls of Pakistani descent. The combined meta-analysis identified 10 genome-wide significant loci. HapMap-2 CEU was used as the imputation panel resulting in a common set of ~2.5 million SNPs across studies. Study-specific GWAS adjusted for age of onset (cases) or age of recruitment (controls), gender, and genetic PCs. The distribution of PGS for T2D in ELSA is depicted in **Figure 19**. The PGS contain 761,488 SNPs that overlapped between the ELSA genetic database and the DIAGRAM GWAS summary statistics; these SNPs were included in PGS for T2D.

Figure 19. Distribution of PGS for Type II Diabetes



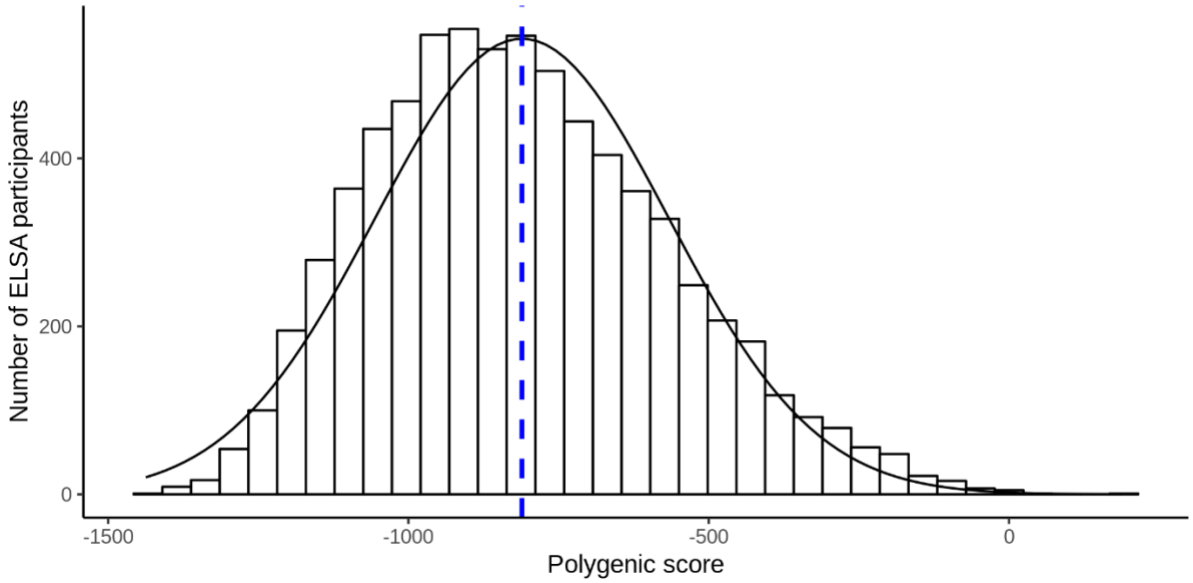
3.3.5.4. Type II Diabetes (2018)

PGSs for Type II Diabetes (T2D, 2018) were created using GWAS meta-analysis that combined the DIAGRAMv3 (stage 1) GWAS meta-analysis with a stage 2 meta-analysis comprising 22669 cases and 58,190 controls genotyped with Metabochip, including 1,178 cases and 2472 controls of Pakistani descent (Pakistan Risk Of Myocardial Infarction Study (PROMIS))[50]. Combining stage 1 and stage 2 meta-analyses included 34840 cases and 114,981 controls overwhelmingly of European descent leading to identification to eight new T2D susceptibility loci at genome-wide significance ($P < 5 \times 10^{-8}$).

3.3.5.5. Rheumatoid Arthritis

The PGSs for Rheumatoid Arthritis (RA) were created using results from a 2014 GWAS that was performed in a total of >100,000 subjects of European and Asian ancestries (29,880 RA cases and 73,758 controls), by evaluating 10 million SNPs. From these analyses, 42 novel RA risk loci at a genome-wide level of significance were discovered, bringing the total to 101 [51]. After applying quality control criteria, whole-genome genotype imputation was performed using 1000 Genomes Project Phase I (α) European (*n*=381) and Asian (*n*=286) data as references. Associations of SNPs with RA were evaluated by logistic regression models assuming additive effects of the allele dosages including top 5 or 10 principal components as covariates (if available) using mach2dat v.1.0.16. To calculate the PGS for RA, the negative ORs value from the GWAS summary statistics (the OR <1), the OR measures were not converted to positive values and the reference allele were flipped to represent phenotype-increasing PGS. The distribution of PGS for RA in ELSA is depicted in **Figure 20**. A total of 8,747,962 SNPs were included in the meta-analysis summary statistics for RA. Of these, 1,100,616 SNPs overlapped with the ELSA genetic database and were included in the PGS for the Rheumatoid arthritis phenotype.

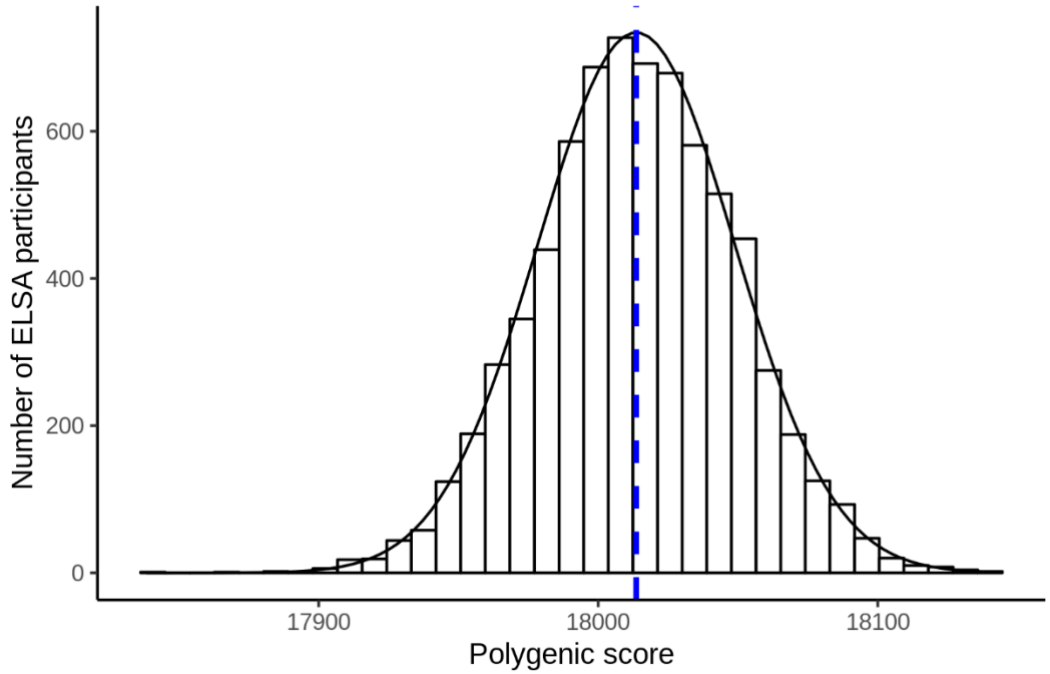
Figure 20. Distribution of PGS for Rheumatoid Arthritis



3.3.5.6. Myocardial Infarction

The PGSs for myocardial infarction (MI) were created using 2015 results from a subgroup analysis of coronary artery disease (CAD) conducted by the Coronary ARtery DIsease Genome wide Replication and Meta-analysis (CARDIoGRAM) consortium[52]. The GWAS is a meta-analysis of 48 studies of mainly European, South Asian, and East Asian, descent imputed using the 1000 Genomes phase 1 v3 training set with 38 million variants. The study interrogated 9.4 million variants and involved 60,801 CAD cases and 123,504 controls. Case status was defined by an inclusive CAD diagnosis (for example, myocardial infarction, acute coronary syndrome, chronic stable angina or coronary stenosis of >50%). 37 previous loci and 10 new loci achieved genome-wide significance in these analyses. MI subgroup analysis was performed in cases with a reported history of MI (~70% of the total number of cases). No additional loci reached genome-wide significance in the MI analysis. The distribution of PGS for MI in ELSA is depicted in **Figure 21**. The European ancestry PGSs contain 1,299,282 SNPs that overlapped between the ELSA genetic database and the GWAS meta-analysis; these SNPs were included in the PGS.

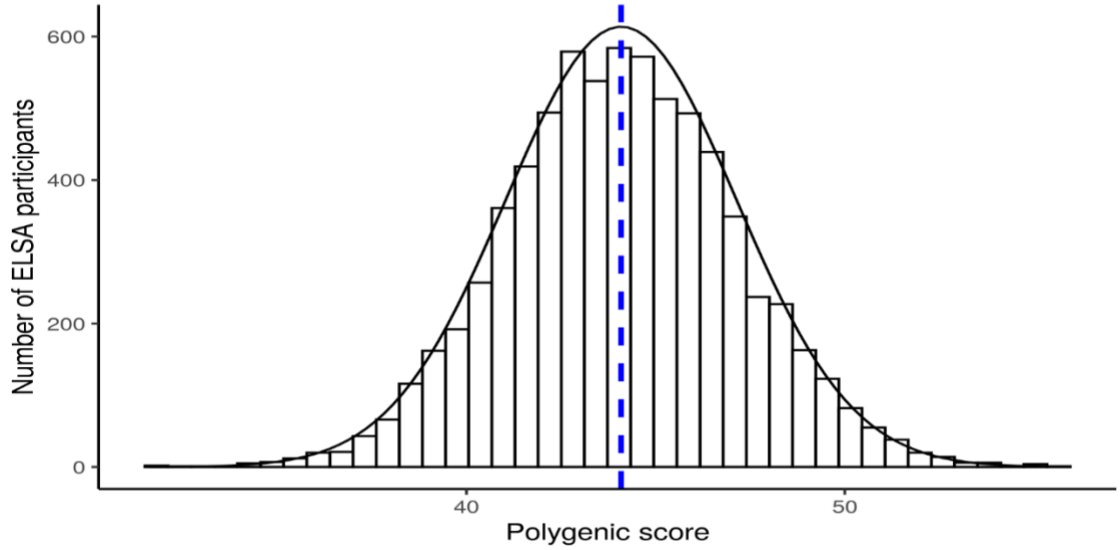
Figure 21. Distribution of PGS for Myocardial Infarction in ELSA



3.3.5.7. Migraine (2016)

PGSs for migraine were calculated using data from a meta-analysis of 22 GWA studies, including data for a total of 59,674 affected subjects and 316,078 controls collected from six tertiary headache clinics and 27 population-based cohorts throughout worldwide collaboration with the International Headache Genetics Consortium (IHGC)[53]. This combined data set contained more than 35,000 new migraine cases not included in previously published GWA studies. These case samples came from both individuals diagnosed by a doctor and individuals with self-reported migraine as stated on questionnaires. The final combined sample consisted of 59,674 case samples and 316,078 controls in 22 non-overlapping case–control data sets. All subjects were of European ancestry (EUR). Missing genotypes were imputed into each sample using a common 1000 Genomes Project reference panel. Association analyses were carried out within each study using logistic regression on the imputed marker dosages, with adjustments made for sex and other covariates where necessary. The association results were combined in an inverse-variance weighted fixed-effects meta-analysis. Markers were filtered for imputation quality and other metrics, leaving 8,094,889 variants for consideration in our primary analysis. It is important to note that only the genetic markers that reached suggestive $P_{GWAS}<5x10^{-6}$ were made available for the public download. Therefore, PGSs for Migraine in ELSA were calculated based on 7208 markers, of which 1188 overlapped with the ELSA data. The distribution of PGS for migraine is depicted in **Figure 22**.

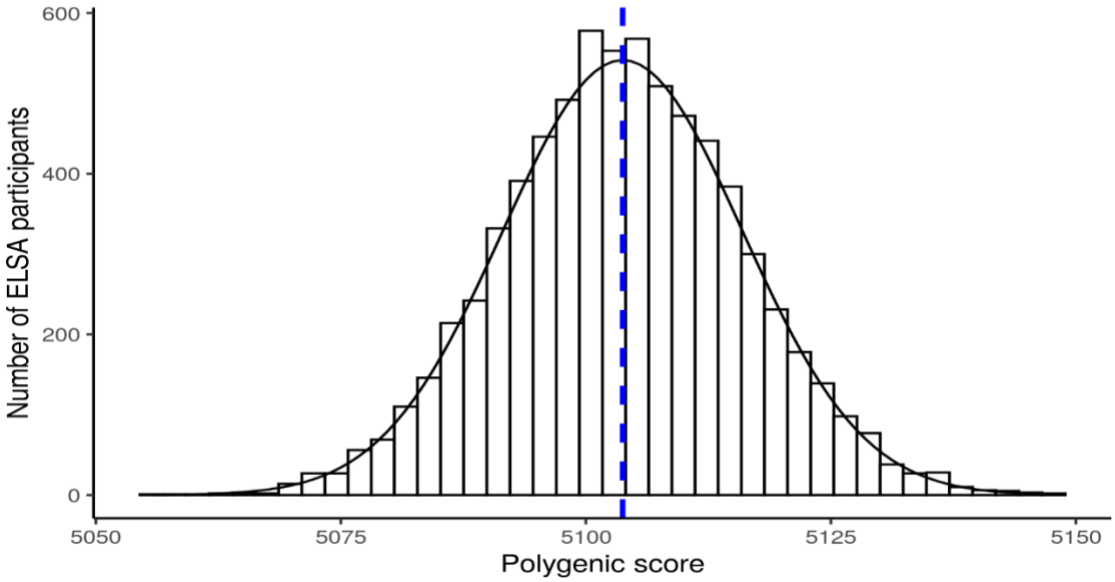
Figure 22. Distribution of PGS for Migraine (2016)



3.3.5.8. Chronic pain

PGSs for chronic pain were calculated using the summary statistics from a large-scale GWAS of Multisite Chronic Pain (MCP) in 387,649 UK Biobank participants[54]. To define MCP phenotype, UK Biobank participants were asked via a touchscreen questionnaire about "pain types experienced in the last month", with possible answers: 'None of the above'; 'Prefer not to answer'; pain at seven different body sites (head, face, neck/shoulder, back, stomach/abdomen, hip, knee); or 'all over the body'. Where patients reported recent pain at one or more body sites, or all over the body, they were additionally asked whether this pain had lasted for 3 months or longer. Those who chose 'all over the body' could not also select from the seven individual body sites. MCP was defined as the sum of body sites at which chronic pain (at least 3 months duration) was recorded: 0 to 7 sites. Those who answered that they had chronic pain 'all over the body' were excluded from the GWAS. The distribution of PGS for chronic pain in ELSA is depicted in **Figure 23**. A total of 1,351,316 SNPs overlapped with the ELSA genetic database with the GWAS summary statistics and were included in the PGS for chronic pain.

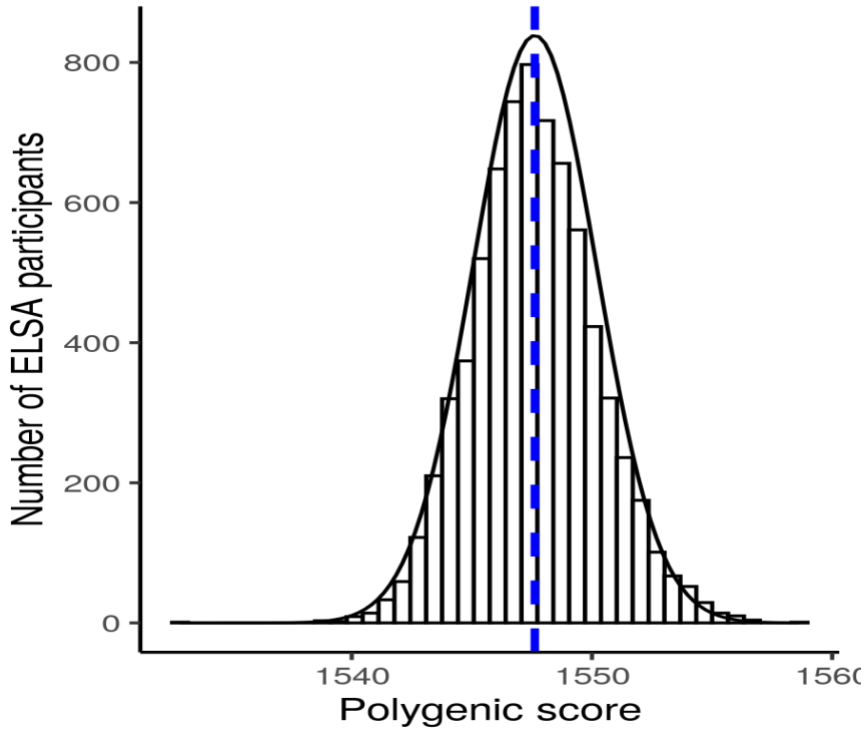
Figure 23. Distribution of PGS for Chronic pain



3.3.5.9. Gait speed

PGSs for Gait speed were calculated using the summary statistic from a meta-analysis of gait speed GWASs in 31478 older adults from 17 cohorts of the CHARGE consortium and validated our results in 2588 older adults from 4 independent studies. Timed walk at usual pace was converted to gait speed (m/s) to harmonize the phenotype across cohorts. The meta-analysis resulted in a list of 536 suggestive genome wide significant SNPs in or near 69 genes. The distribution of PGS for Gait speed in ELSA is depicted in **Figure 24**. A total of 772,690 SNPs overlapped with the ELSA genetic database with the GWAS summary statistics and were included in the PGS for Gait speed phenotype.

Figure 24. Distribution of PGS for Gait speed



3.3.5.10. Relative grip strength

PGSs for Relative grip strength using the summary statistics from a large genome-wide association study to discover genetic variation associated with muscular strength, and to evaluate shared genetic aetiology with and causal effects of muscular strength on several health indicators[55]. Association analysis was conducted with PLINK (version 2.0) assuming an additive model for association. Age, sex, genotype array and 10 principal components were used as covariates. Analysis was restricted to single. In our discovery analysis of 223,315 individuals, we identified 101 loci associated with grip strength ($P_{GWAS} < 5 \times 10^{-8}$). Of these, 64 were associated ($P < 0.01$ and consistent direction) also in the replication dataset (N = 111,610). The distribution of PGS for Relative grip strength in ELSA is depicted in **Figure 25**. A total of 1344065 SNPs overlapped with the ELSA genetic database with the GWAS summary statistics and were included in the PGS for Relative grip strength phenotype.

Figure 25. Distribution of PGS for relative grip strength

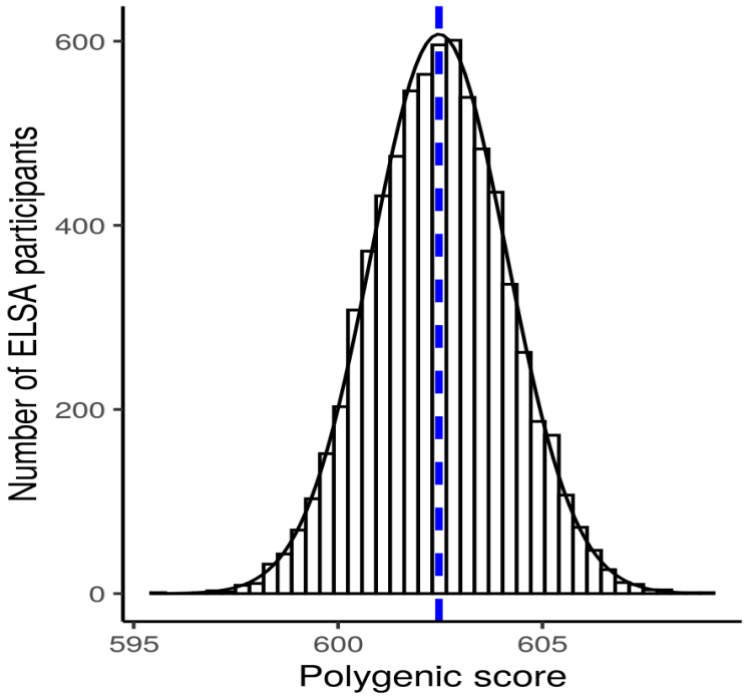


Table 10. The summary statistics for PGS for physical health outcomes

PGS	Sample Size	Minimum	Maximum	Range	Median	Mean	SE (mean)
Coronary Artery Disease (2011)	7183	14234.7	14525.7	291	14392.6	14392.3	0.43
Coronary Artery Disease (2018)	7183	49133.5	49855.4	721.9	49463.6	49463.9	1.09
Chronic pain	7183	5055.8	5147.8	92.0	5103. 7	5103.7	0.14
Gait speed	7183	1532.8	1558.6	25.8	1547.5	1547.6	0.03
Migraine (2016)	7183	31.6	55.5	23.9	44.0	44.1	0.03
Myocardial infarction	7183	17838.4	18137.8	299.4	18013.4	18013.6	0.42
Relative grip strength	7183	595.5	608.9	13.4	602.5	602.5	0.02
Rheumatoid arthritis	7183	-1437.4	187.5	1624.8	-831.7	-811.0	2.87
Type II Diabetes (2012)	7183	16806.4	17179.5	373.1	16998.8	16998.7	0.57
Type II Diabetes (2018)	7183	9589.3	9851.1	261.8	9713.5	9714.2	0.40

PGS, polygenic score; SE, standard error

Table 11. Sources of the GWAS summary statistics used for these physical health outcomes

Phenotype	Consortium	GWAS SNPs	Overlapping with ELSA	GWAS meta-analysis citation	Source of base data
Chronic pain (2018)	UK Biobank	9,926,106	1,351,316	Johnston et al (2019)[54]	https://data.broadinstitute.org/alkesgroup/BOLT-LMM/
Coronary Artery Disease (2011)	CARDIoGRAM	2,420,360	783,413	Schunkert et al. (2011)[47]	www.cardiogramplusc4d.org (cad.add.160614.website.txt)
Coronary Artery Disease (2018)	UK Biobank	9,087,334	1,349,976	Fall et al (2018)[48]	https://grasp.nhlbi.nih.gov/FullResults.aspx
Gait speed (2017)	CHARGE	2,474,503	772,690	Ben-Avraham et al (2017)[56]	https://grasp.nhlbi.nih.gov/FullResults.aspx
Migraine (any)	-	7,208	1,188	Gormley et al (2016)[53]	http://www.headachegenetics.org/content/datasets-and-cohorts
Myocardial infarction	CARDIoGRAM	9,289,491	1,299,282	CARDIoGRAMplusC4D Consortium. (2015)[52]	www.cardiogramplusc4d.org (mi.add.030315.website.txt)
Relative hand grip	UK Biobank	15,544,142	1,344,065	Tikkanen et al (2018)[55]	https://grasp.nhlbi.nih.gov/FullResults.aspx
Rheumatoid arthritis	-	8,747,962	1,100,616	Okada et al. (2014)[51]	http://plaza.umin.ac.jp/~yokada/datasource/software.htm
Type II Diabetes (2012)	DIAGRAM	2,473,441	761,488	Morris et al. (2012)[49]	https://grasp.nhlbi.nih.gov/FullResults.aspx
Type II Diabetes (2018)	DIAGRAMv3	5,053,015	920,194	Xue et al (2018)[50]	http://www.diagram-consortium.org/downloads.html (DIAGRAMv3.2012DEC17.txt).

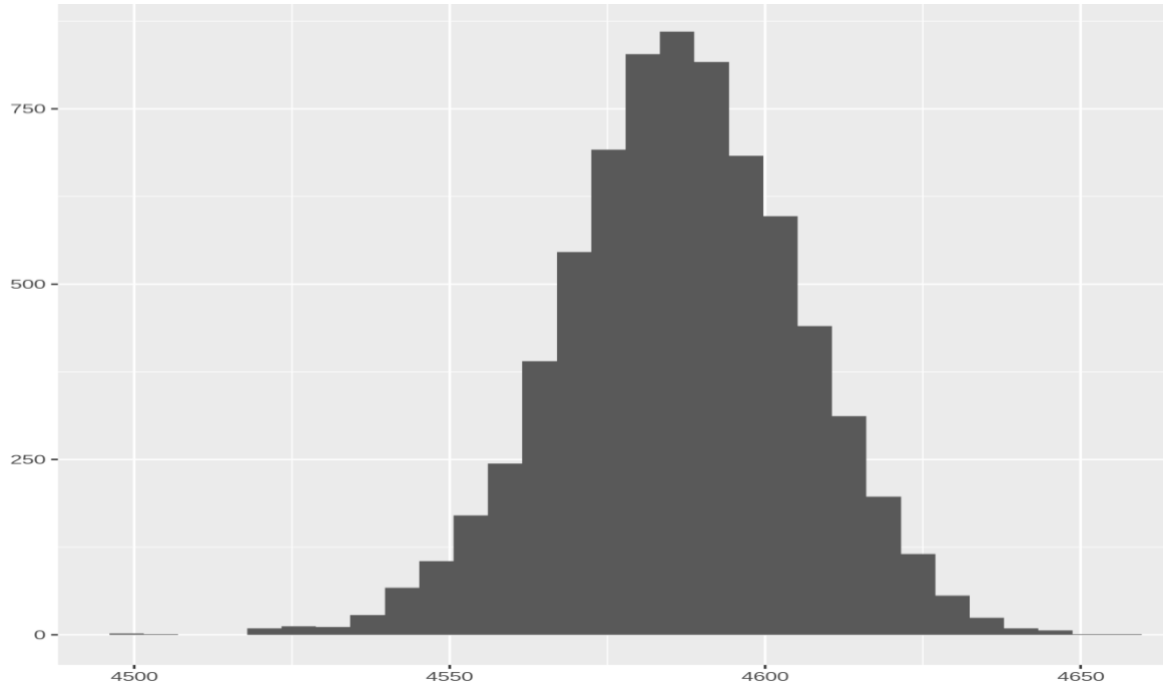
CARDIoGRAM, Coronary ARtery DIsease Genome wide Replication and Meta-analysis; CHARGE, Heart and Aging Research in Genomic Epidemiology consortium; DIAGRAM, DIAbetes Genetics Replication and Meta-analysis Consortium

3.3.6. ANTHROPOMORPHIC TRAITS

3.3.6.1. Height

PGS for height was created using the results from a 2014 study conducted by the Genetic Investigation of ANthropometric Traits (GIANT) consortium [57]. The GIANT height meta-analysis included 253,288 individuals from 79 studies imputed to HapMap II with a total of 2,550,858 SNPs. Replication was performed in a sample of 80,067 individuals. The participating studies adjusted for age and genetic PCs in their GWASs. Height was measured as sex standardised height (in centimetres). There were 697 GWAS significant SNPs identified that together explain one-fifth of heritability for adult height. The distribution of PGS for Height in ELSA is depicted in **Figure 26**. The PGS contains 831,045 SNPs that overlapped between the ELSA genetic database and the GIANT GWAS meta-analysis and that were included PGS for this phenotype.

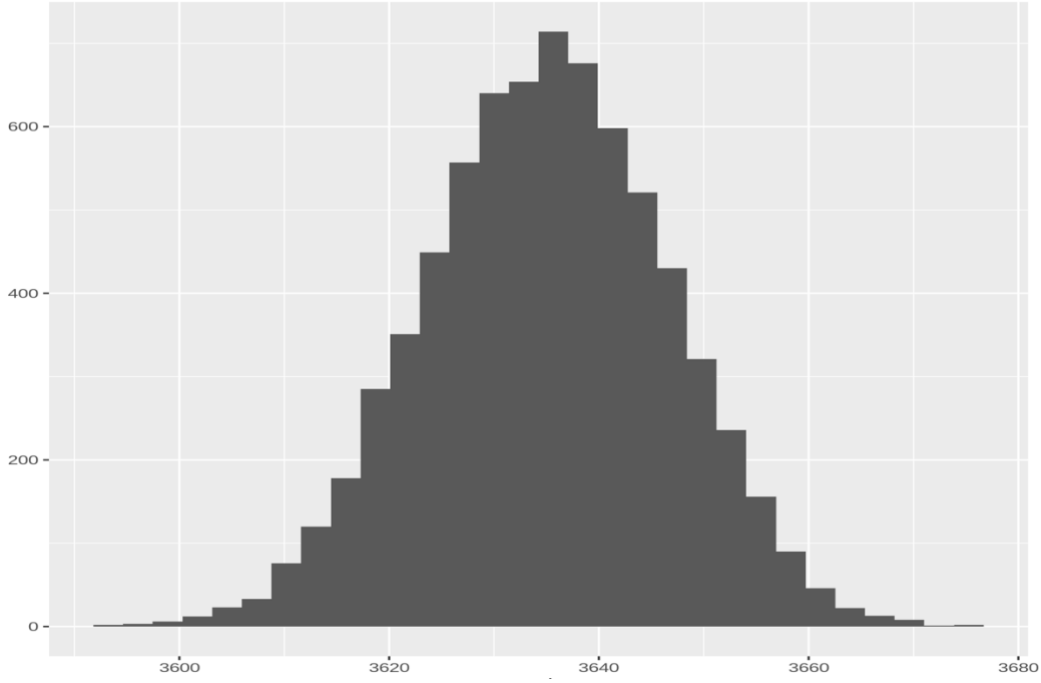
Figure 26. Distribution of PGS for Height



3.3.6.2. Body Mass Index (BMI) - 2015

PGS for BMI was created using results from a 2015 GWAS conducted by the Genetic Investigation of ANthropometric Traits (GIANT) consortium [58]. The GIANT GWAS meta-analysis was performed on a sample of 234,069 individuals from 80 studies across 2,550,021 SNPs, and separately in a Metabochip (MC) meta-analysis on a sample of 88,137 individuals from 34 studies across 156,997 SNPs. The joint GWAS and MC meta-analysis comprised of 322,154 individuals of European descent and 17,072 individuals of non-European descent identified 97 GWS loci associated with BMI, 56 of which were novel. These loci accounted for 2.7% of the variation in BMI and suggest that as much as 21% of BMI variation can be accounted for by common genetic variation. Adjustment covariates within each contributing cohort GWAS included age, age^2, sex and genetic PCs. The distribution of PGS for BMI in ELSA is depicted in **Figure 27**. The PGS contains 795,650 SNPs that overlapped between the ELSA genetic database and the GIANT GWAS meta-analysis which were included in PGS.

Figure 27. Distribution of PGS for BMI (2015)

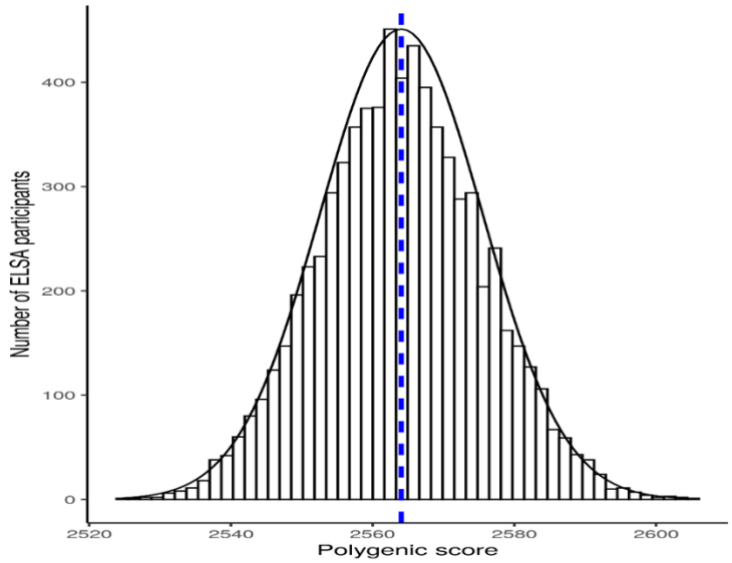


3.3.6.3. Body Mass Index (BMI) - 2018

PGS for BMI was created using results from the present study which was part of a larger effort led by the GIANT consortium (2018)[59]. A GWAS of BMI was run in 456,426 UKB participants using linear mixed model association testing implemented in BOLT-LMM v2.3 software assuming an infinitesimal model. Altogether, used 711,933 HM3 SNPs (LD pruned for SNPs with r^2>0.9) were used in the analyses. The model was adjusted for age, sex, recruitment centre, genotyping batches and 10 PCs

calculated from 132,102 out of the 147,604 genotyped SNPs pre-selected by the UKB quality control team for PC analysis. The distribution of PGS for BMI in ELSA is depicted in **Figure 28**. A total of 2529253 SNPs were included in the summary statistics. Of these, 798737 SNPs overlapped with the ELSA genetic database and were included in the PGS for BMI (2018).

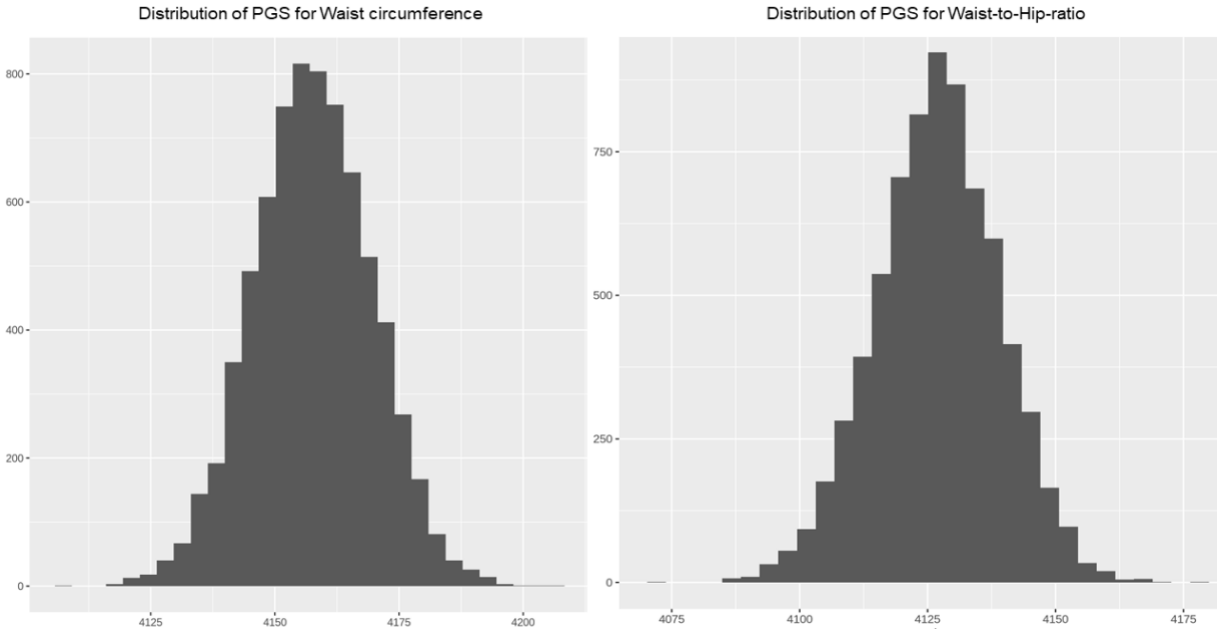
Figure 28. Distribution of PGS for BMI (2018)



3.3.6.4. Waist circumference & Waist-Hip Ratio

PGS for waist circumference (WC) and waist-to-hip ratio (WHR) were created using results from a 2015 study conducted by the Genetic Investigation of ANthropometric Traits (GIANT) consortium [60]. GWAS meta-analysis was performed on a sample of 142762 individuals from 57 studies, and separately in a Metabochip (MC) meta-analysis on a sample of 67,326 individuals from 44 studies across 124,196 SNPs. A joint GWAS and MC meta-analysis was then carried out on 210,088 individuals across 93057 SNPs. The GWAS identified 49 loci associated with WHR and an additional 19 loci associated with WC at the genome-wide significance level. Association analyses adjusted for age, age^2, study-specific covariates if necessary, and BMI. The distributions of PGSs for Waist circumference & Waist-Hip Ratio are depicted in **Figure 29**. PGS for WC in ELSA contain 801,114 SNPs that overlapped between the ELSA genetic database and the GIANT GWAS meta-analysis. PGS for WHR in ELSA contains 801,207 SNPs that overlapped between the ELSA genetic database and the GIANT GWAS meta-analysis.

Figure 29. Distribution of PGS for WC and WHR



PGS, polygenic score; WC, waist circumference; WHR, Waist-Hip Ratio

Table 12. Descriptive statistics for PGS for anthropomorphic traits

PGS	Sample Size	Minimum	Maximum	Range	Median	Mean	SE (mean)
BMI (2015)	7183	3594.0	3676	82.1	3635.1	3635	0.14
BMI (2018)	7183	2524.6	2605.1	80.5	2563.9	2564.0	0.13
Height	7183	4498.4	4656.5	158.1	4586.6	4586.5	0.22
Waist circumference	7183	4106.3	4205.5	99.2	4157.5	4157.6	0.14
Waist-to-hip ratio	7183	4070.5	4176.6	106.1	4127.1	4126.9	0.14

PGS, polygenic score; SE, standard error; WC, waist circumference; WHR, Waist-Hip Ratio

Table 13. Sources of the GWAS summary statistics used for these anthropomorphic traits

Phenotype	Consortium	GWAS SNPs	Overlapping with ELSA	GWAS meta-analysis citation	Source of base data
Body Mass Index (2015)	GIANT	2,554,623	795,650	Locke et al. (2015)[58]	https://www.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium_data_files
Body Mass Index (2018)	GIANT	2,529,253	798,737	Yengo et al (2018)[59]	https://portals.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium_data_files
Height	GIANT	2,550,858	831,045	Wood et al. (2014)[57]	https://www.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium_data_files
Waist circumference	GIANT	2,565,407	801,114	Shungin et al. (2015)[60]	https://www.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium_data_files; WC: GIANT 2015 WC COMBINED EUR.txt.gz
Waist-to-hip ratio	GIANT	2,542,431	801,207		https://www.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium_data_files; WHR: GIANT 2015 WHR COMBINED EUR.txt.gz

3.3.7. BEHAVIOURAL TRAITS

3.3.7.1. SMOKING BEHAVIOUR

3.3.7.1.1. Smoking ever (2010)

PGS for smoking behaviours in ELSA was contracted using the results from the Tobacco and Genetics (TAG) Consortium (2010) [61]. The TAG examined four smoking phenotypes - smoking initiation (ever versus never been a regular smoker), age of smoking initiation, smoking quantity (number of cigarettes smoked per day, CPD) and smoking cessation (former versus current smokers) - among people of European ancestry. In ELSA we created PGSs for two of these smoking phenotypes - 1) CPD and 2) Smoking initiation. The TAG GWAS included a total of 74,053 participants in the discovery phase of the analysis; another 73,853 participants were included in a follow-up meta-analysis of the 15 most significant regions. The included studies were genotyped on six different platforms. Genotype imputations resulted in a common set of ~2.5 million of SNPs.

3.3.7.1.2 Number of cigarettes smoked per day - 2010

In the TAG consortium Number of cigarettes smoked per day (CPD) was measured as either average CPD or maximum CPD in a sample of 73,853 individuals. Study-specific GWAS controlled for imputed allele dosage for a SNP plus whether a subject was classified as a case in the primary study. If the primary study was a case-control design and the phenotype being studied was known to be associated with smoking, the GWAS adjusted for case status to reduce the potential confounding. Analyses were run and meta-analysed separately for males and females. The distribution of PGS for CPD in ELSA is depicted in **Figure 39**; the summary statistics for PGS for CPD are provided in **Table 37**. TAG GWAS summary statistics contained 2,459,118 SNPs of which 803,092 SNPs overlapped with the ELSA genetic database and were included in the PGS for the CPD phenotype.

3.3.7.1.3. Smoking initiation (ever/never) - 2010

In the TAG consortium individuals who were recorded as having ever been regular smokers were defined as those who reported having smoked at least 100 cigarettes during their lifetime, and never regular smokers were defined as those who reported having smoked between 0 and 99 cigarettes during their lifetime. Study-specific GWASs controlled for imputed allele dosage for a SNP plus whether a subject was classified as a case in the primary study. If the primary study was a case-control in design and the phenotype being studied was known to be associated with smoking, the GWAS adjusted for case status to reduce potential confounding. Analyses were run and meta-analysed separately for males and females. The distribution of PGS for Smoking initiation in ELSA is depicted in **Figure 39**; the summary statistics for PGS for smoking initiation are provided in **Table 37**. The TAG GWAS summary statistics for this smoking phenotype was based on the sample of 143,023 individuals and contained 2,455,846 SNPs; of these, 804,337 SNPs overlapped with the ELSA genetic database and were included in the PGS for smoking initiation phenotype.

3.3.7.1.4. Age of smoking initiation (2019)

The PGSs for the age of smoking initiation (2019) was calculated using summary statistics from the study that combined study-level summary association data from 1.2 million individuals

of European ancestry[62]. These GWAS summary statistics are publicly available; the link to the website and the file can be found in **Error! Reference source not found.** The majority of studies imputed their genotypes to the Haplotype Reference Consortium using the University of Michigan Imputation Server. All studies used either Minimac3or IMPUTE2 for imputation. Sample sizes ranged from 337,334 for cigarettes per day to 1,232,091 for smoking initiation. All studies adjusted each trait for age, age squared, sex, and genetic principal components and other study specific covariates in their analyses (e.g., case-control status, site in multi-site studies). Age of smoking initiation was defined as age at which an individual started smoking cigarettes regularly. The GWAS summary statistics contained 11,802,365 SNPs of which 1,339,648 SNPs overlapped with the ELSA genetic database and were included in the PGS for Age of smoking initiation.

3.3.7.1.5. Smoking Cessation (2019)

The PGSs for Smoking Cessation (2019) was calculated using summary statistics from[62]. Smoking cessation phenotype was defined as a binary trait: former vs current smokers. The GWAS summary statistics contained 12,197,133 SNPs of which 1,347,877 SNPs overlapped with the ELSA genetic database and were included in the PGS for Smoking Cessation (2019).

3.3.7.1.6. Smoking initiation (2019)

The PGSs for the smoking initiation (2019) was calculated using summary statistics from the study that combined study-level summary association data from up to 59 studies of European ancestry[62]. For more details please refer to section 3.3.7.1.2. Smoking initiation (binary trait: ever vs never smokers) was defined as individuals who have smoked >99 cigarettes in their lifetime, which is consistent with the definition by the Centre for Disease Control. The distribution of PGS for Smoking initiation (2019). The GWAS summary statistics contained 11,916,706 SNPs of which 1,341,672 SNPs overlapped with the ELSA genetic database and were included in the PGS for Smoking initiation (2019).

3.3.7.1.7. Number of cigarettes per day (2019)

The PGSs for Number of cigarettes per day (2019) was calculated using summary statistics from genome-wide association study of 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use[62]. For more details please refer to section 3.3.7.1.2. Number of cigarettes per day was defined as the average number of cigarettes smoked per day, either as a current smoker or former smoker, and whether self-rolled or manufactured are smoked (most studies did not distinguish). Individuals who either never smoked, or for whom there is no available data (e.g., someone was a former smoker, but for whom former smoking was never assessed) were set to missing. The distribution of PGS for The GWAS summary statistics contained 12,003,613 SNPs of which 1,341,672 SNPs overlapped with the ELSA genetic database and were included in the PGS for Number of cigarettes per day (2019).

3.3.7.2. ALCOHOL INTAKE

3.3.7.2.1. Daily Alcohol Intake (2016)

PGS for smoking behaviours in ELSA was calculated using the results from the genome-wide association meta-analysis and replication study among >105,000 individuals of European

ancestry[63]. These GWAS summary statistics are publicly available. Alcohol intake in grams of alcohol per day was estimated by each cohort based on information about drinking frequency and type of alcohol consumed. The grams per day variable was then $\log_{10}$ transformed before the analysis. Sex-specific residuals were derived by regressing alcohol in $\log_{10}$ (grams per day) in a linear model on age, age^2, weight, and if applicable, study site and principal components to account for population structure. The sex-specific residuals were pooled and used as the main phenotype for subsequent analyses. The GWAS summary statistics for the Daily Alcohol Intake phenotype included 2,462,742 SNPs; of these, 800,524 SNPs overlapped with the ELSA genetic database and were included in the PGS for Daily Alcohol Intake phenotype.

3.3.7.2.2. Drinking alcohol per week (2019)

The PGSs for the Drinking alcohol per week (2019) was calculated using summary statistics from genome-wide association study of 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use[62]. The GWAS summary statistics for the Daily Alcohol Intake phenotype included 11916706 SNPs; of these, 1342722 SNPs overlapped with the ELSA genetic database and were included in the PGS for Daily Alcohol Intake phenotype.

Table 14. The summary statistics for PGS for behavioural traits

PGS	Sample Size	Minimum	Maximum	Range	Median	Mean	SE (mean)
Age of smoking initiation (2019)	7183	5319.9	5273.3	5273.1	5236.3	83.7	0.14
Daily alcohol intake (2019)	7183	2603.6	2666.0	62.4	2642.5	2642.4	0.08
Drinking alcohol per week (2019)	7183	3788.4	3856.3	67.9	3819.7	3819.7	0.10
Number of cigarettes smoked daily (2019)	7183	10793.5	10716.4	10716.2	10625.0	168.5	0.28
Smoking cessation (2019)	7183	9744.9	9925.2	180.3	9827.1	9828.0	0.28
Smoking initiation (2010)	7183	13540.9	13415.2	13414.7	13289	251.9	0.39
Smoking initiation (2019)	7183	7554.5	7468.2	7467.8	7387.4	167.1	0.25
Smoking initiation (2019)	7183	10625.0	10793.5	168.5	10716.2	10716.3	0.28
Number of cigarettes smoked (2010)	7183	95447.3	94694.4	94696.9	93951.2	1496.1	2.50

PGS, polygenic score; SE, standard error

Table 15. Outlines details of the GWAS summary statistics used for these behavioural traits

Phenotype	Consortium	GWAS SNPs	Overlapping with ELSA	GWAS meta-analysis citation	Source of base data
Smoking Initiation (ever/never) (2010)	TAG	2,455,846	804,337	Tobacco and Genetics Consortium (2010) [61]	https://www.med.unc.edu/pgc/results-and-downloads (tag.evrsmk.tbl.gz)
Number of cigarettes per day (2010)	TAG	2,459,118	803,092		https://www.med.unc.edu/pgc/results-and-downloads (tag.cpd.tbl.gz)
Age of smoking initiation (2019)	GSCAN	11,802,365	1,339,648	Liu et al (2019)[62]	https://genome.psych.umn.edu/index.php/GSCAN
Smoking Cessation (2019)		12,197,133	1,347,877		
Smoking initiation (2019)		11,916,706	1,341,672		
Number of cigarettes per day (2019)					
Daily Alcohol Intake (2016)	-	2,462,742	800,524	Schumann et al (2016)[63]	https://grasp.nhlbi.nih.gov/FullResults.aspx
Drinking alcohol per week (2019)	GSCAN	12,003,613	1,341,672	Liu et al (2019)[62]	https://genome.psych.umn.edu/index.php/GSCAN

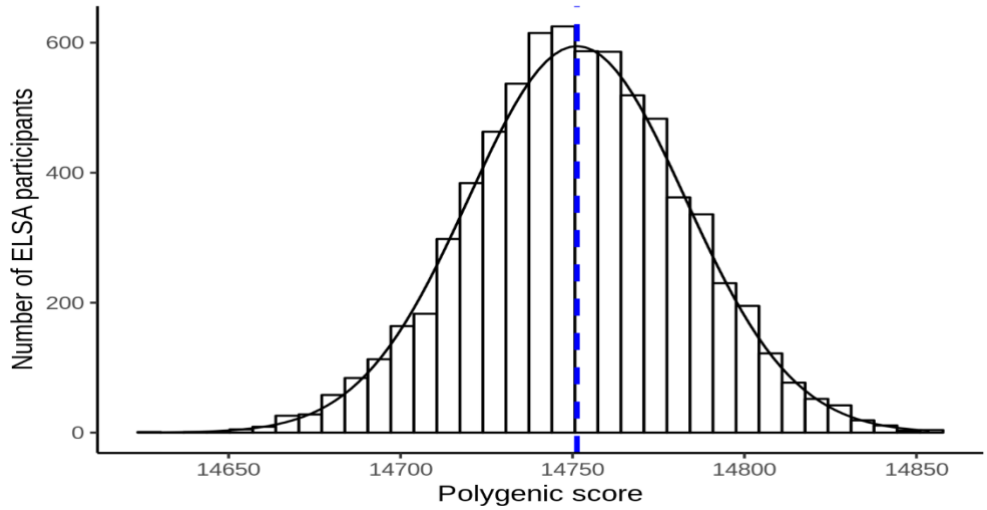
TAG, Tobacco and Genetics; GSCAN, GWAS & Sequencing Consortium of Alcohol and Nicotine use

3.3.8. BIOLOGICAL OUTCOMES

3.3.8.1. Morning Plasma cortisol

PGS for morning plasma cortisol in ELSA was contracted using the results from the CORtisol NETwork (CORNET) consortium which undertook the GWAS meta-analysis for plasma cortisol in 12,597 Caucasian participants from 11 western European population-based cohorts, and replicated their results in 2,795 participants from three independent cohorts [64]. Cortisol was measured by immunoassay in blood samples collected from study participants between 07:00h and 11:00h. Each study performed single marker association tests, and study-specific linear regression models which used z-scores of log-transformed cortisol, additive SNP effects, and were adjusted for age and sex (model 1); age, sex, and smoking (model 2); or age, sex, smoking and body mass index (model 3). Imputation of the gene-chip results used the HapMap CEU population, build 36. The results indicate that <1% of variance in plasma cortisol is accounted for by genetic variation in a single region of chromosome 14. The CORNET GWAS summary statistics for this phenotype contained 2,660,191 SNPs; of these, 837,709 SNPs overlapped with the ELSA genetic database and were included in the PGS for Morning Plasma Cortisol phenotype.

Figure 30. Distribution of PGS for Morning Plasma cortisol

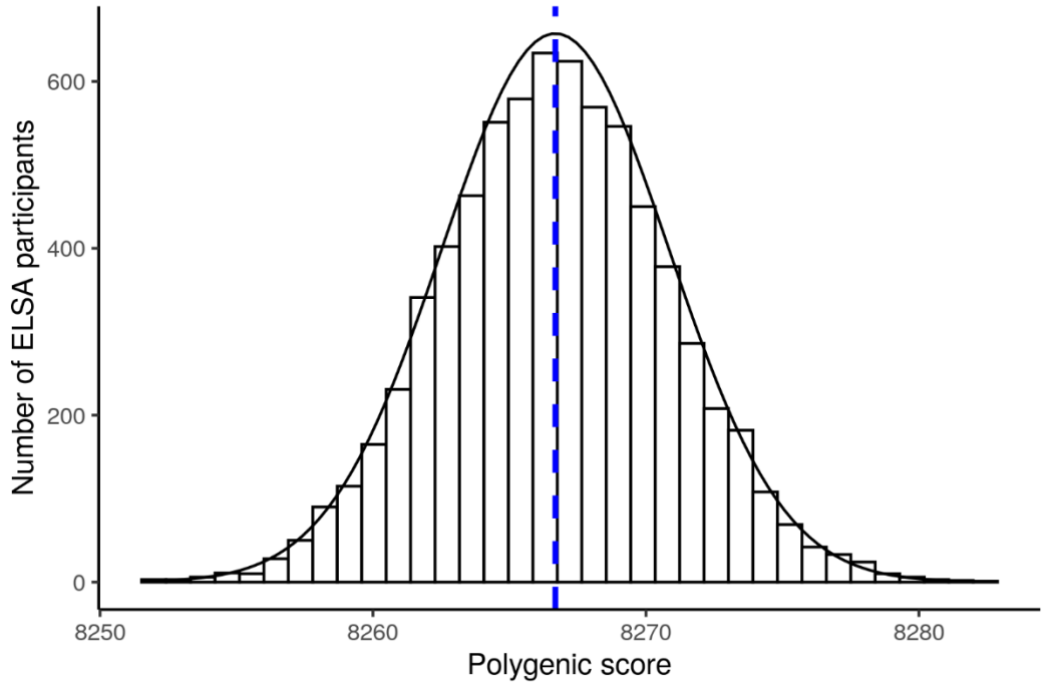


3.3.8.2. C-reactive protein (2018)

PGS for C-reactive protein in ELSA was contracted using the results from two genome-wide association studies (GWASs), based on both HapMap and 1000 Genomes imputed data and encompassing data from 88 studies comprising 204,402 European individuals[65]. The GWAS meta-analyses of C-reactive protein revealed 58 distinct genetic loci ($p < 5 \times 10^{-8}$). After adjustment for body mass index in the regression analysis, the associations at all except three loci remained. The lead

variants at the distinct loci explained up to 7.0% of the variance in circulating amounts of C-reactive protein. Further 66 gene sets that were organized in two substantially correlated clusters were identified, one mainly composed of immune pathways and the other characterized by metabolic pathways in the liver. The summary statistics for C-reactive protein were obtained on request from the authors. The distribution of PGS for C-reactive protein in ELSA is depicted in **Figure 31**. The GWAS summary statistics for this phenotype contained 10019203 SNPs; of these, 1,301,076 SNPs overlapped with the ELSA genetic database and were included in the PGS for C-reactive protein phenotype.

Figure 31. Distribution of PGSs for C-reactive protein



3.3.8.3. C-reactive protein (2022)

After exclusions, 427367 UKB participants contributed to the GWAS analysis which identified 49164 SNPs associated with CRP levels (at genome wide significance (GWS) of $p < 5 \times 10^{-8}$) [66]. Linear Mixed Model (LMM) regression using BOLT-LMM version 2.343 was performed on CRP levels in UKB. This model accounts for cryptic relatedness within the sample. Here, an additive genetic model was used for all 8.9 million measured and imputed genetic variants. The model was adjusted for age, sex, UKB array (UKB vs UK BiLEVE to account for the different genotyping chips) and 40 genetic principal components. Further, serum CRP levels (mg/l was measured by immunoturbidimetry). CRP levels were transformed using natural log and the resulting range included was from −2.53 to 4.38, excluding individuals with extreme values ±4 SD from the mean. Individuals on immune modulating drugs, with auto-immune related diseases/disorders, which constituted 1.8% of the sample, were removed.

3.3.8.4. Blood traits

Blood cell counts and indices are quantitative clinical laboratory measures that reflect hematopoietic progenitor cell production, hemoglobin synthesis, maturation, release from the bone marrow, and clearance of mature or senescent blood cells from the circulation. This GWAS meta-analyses included data from UK Biobank and a large-scale international collaborative effort, including data for 563,085 European ancestry participants, and discover 5,106 new genetic variants independently associated with 29 blood cell phenotypes covering a range of variation impacting hematopoiesis [67, 68].

The following phenotypes were examined:

Basophil count (BASO) (percentage of white cells that are basophils × WBC) 109/L Relative basophil count = percentage of white cells that are basophils × WBC

Eosinophil count (EOS) (percentage of white cells that are eosinophils × WBC) 109/L Relative eosinophil count = percentage of white cells that are eosinophils × WBC

Hematocrit (HCT) % Proportion of blood made up of red blood cells

Hemoglobin (HGB) concentration g/dL Concentration of hemoglobin in each volume of blood

Lymphocyte count (LYM) (percentage of white cells that are lymphocytes × WBC) 109/L: Relative lymphocyte count = percentage of white cells that are lymphocytes × WBC

Mean corpuscular hemoglobin (MCH) pg: Average mass of hemoglobin per red blood cell

Mean corpuscular volume (MCV) fL: Average volume of red blood cells

Mean platelet volume (MPV) fL: Average volume of platelets

Neutrophil count (NEU) (percentage of white cells that are neutrophils × WBC) 109/L Relative neutrophil count = percentage of white cells that are neutrophils × WBC

Platelet count (PLT) 109/L: Number of platelets per unit volume of blood

Red blood cell (RBC) count 1012/L: Number of red blood cells per unit volume of blood

Red cell distribution (RDW) width % Range of variation of red blood cell volume

White blood cell (WBC) count 109/L: Aggregate number of white blood cells per unit volume of blood

Raw phenotypes were regressed on age, age- squared, sex, principal components and cohort specific covariates (e.g., study center, cohort, etc) if needed, WBC related traits were $\log_{10}$ transformed before regression modeling. Residuals from the modeling were obtained and then inverse normalized for cohort level association analysis or GWAS. The cohort level association analyses were then conducted using a linear mixed effects model to account for known or cryptic relatedness (e.g., BOLT-LMM, EPACTS https://github.com/statgen/EPACTS and rvtests with the additive genetic model. Linear mixed effects models have been shown to effectively account for both population structure and inter-individual relatedness within the UK Biobank cohort, along with having increased discovery power over simple linear regression with principal components.

Figure 32. represents correlations between each quantitative clinical laboratory measures

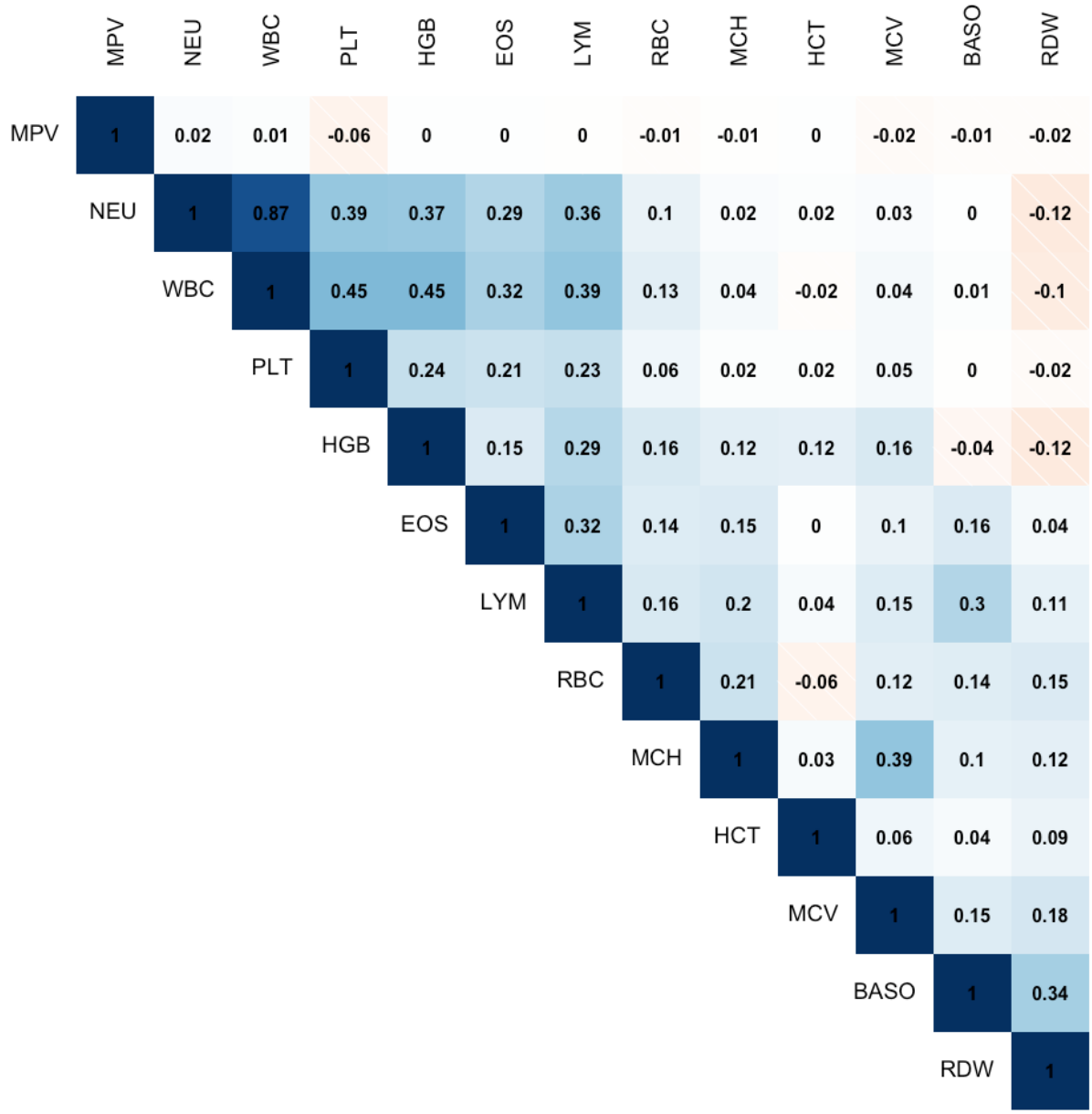


Table 16. The summary statistics for PGS for biological outcomes

Biological outcomes	Sample Size	Minimum	Maximum	Range	Median	Mean	SE (mean)
Morning plasma cortisol	7183	14626.3	14853.8	227.5	14750.8	14751.3	0.37
C-reactive protein (2018)	7183	8251.9	8282.4	30.5	8266.7	8266.7	0.05
C-reactive protein (2022)	7183	4583.5	4692.0	108.5	4636.8	4636.9	0.18
Eosinophil count (EOS)	7183	6838.7	7795.1	956.4	7374.6	7372.0	1.28
Hematocrit (HCT)	7183	1537.0	1615.6	78.6	1579.58	1579.5	0.13
Hemoglobin (HGB)	7183	5563.0	5919.1	356.1	5769.3	5766.7	0.55
Lymphocyte count (LYM)	7183	6969.4	7801.7	832.3	7461.2	7456.3	1.30
Mean corpuscular hemoglobin (MCH)	7183	7869.0	9181.1	1312.1	8561.6	8558.6	1.98
Mean corpuscular volume (MCV)	7183	7287.1	8412.8	1125.7	7810.1	7809.9	1.51
Mean platelet volume (MPV)	7183	34548.1	34867.7	319.6	34705.8	34705.6	0.50
Basophil count (BASO)	7183	2471.5	3324.3	852.8	2981.0	29798.0	1.30
Neutrophil count (NEU)	7183	4999.0	5276.83	280.87	5166.4	5160.6	0.47
Platelet count (PLT)	7183	5547.13	5805.73	258.6	5681.0	5681.0	0.38
Red blood cell (RBC)	7183	7581.76	8629.73	1048.0	8100.54	8099.6	1.74
Red cell distribution (RDW)	7183	7595.6	8666.0	1070.4	8099.2	8100.1	1.82
White blood cell (WBC)	7183	549.1	5858.8	419.6	5701.0	5692.0	0.68

PGS, polygenic score; SE, standard error

Table 17. Sources of the GWAS summary statistics used for these biological outcomes

Phenotype	Consortium	GWAS SNPs	Overlapping with ELSA	GWAS meta-analysis citation	Source of base data
Plasma cortisol (morning)	CORNET	2,660,191	837,709	Bolton et al. (2014)[64]	https://datashare.is.ed.ac.uk/handle/10283/2787
C-reactive protein (2018)	CHARGE	10,019,203	1,301,076	Ligthart et al (2018)[65]	On request from the authors
C-reactive protein (2022)	UKB and CHARGE	9,967,405	1,331,207	Said et al (2022)[66]	https://www.ebi.ac.uk/gwas/studies/GCST90029070
Eosinophil count (2020)	UK Biobank	26,776,137	1,357,065	Vuckovich et al (2020) [67]	https://grasp.nhlbi.nih.gov/FullResults.aspx
Hematocrit (2020)		27,103,386	400,380		
Hemoglobin (2020)		27,102,009	1,361,427		
Lymphocyte count (2020)		26,949,845	1,349,070		
Mean corpuscular hemoglobin (2020)		26,861,301	1,358,545		
Mean corpuscular volume (2020)		27,093,177	1,355,090		
Mean platelet volume (2020)		22,594,326	1,355,090		
Basophil count (2020)		26,797,909	545,506		
Neutrophil count (2020)		26,745,219	1,352,901		
Platelet count (2020)		26,985,305	1,352,668		
Red blood cell (2020)		27,099,926	1,357,201		
Red cell distribution (2020)		26,978,706	1,351,605		
White blood cell (2020)		27,090,932	1,352,140		

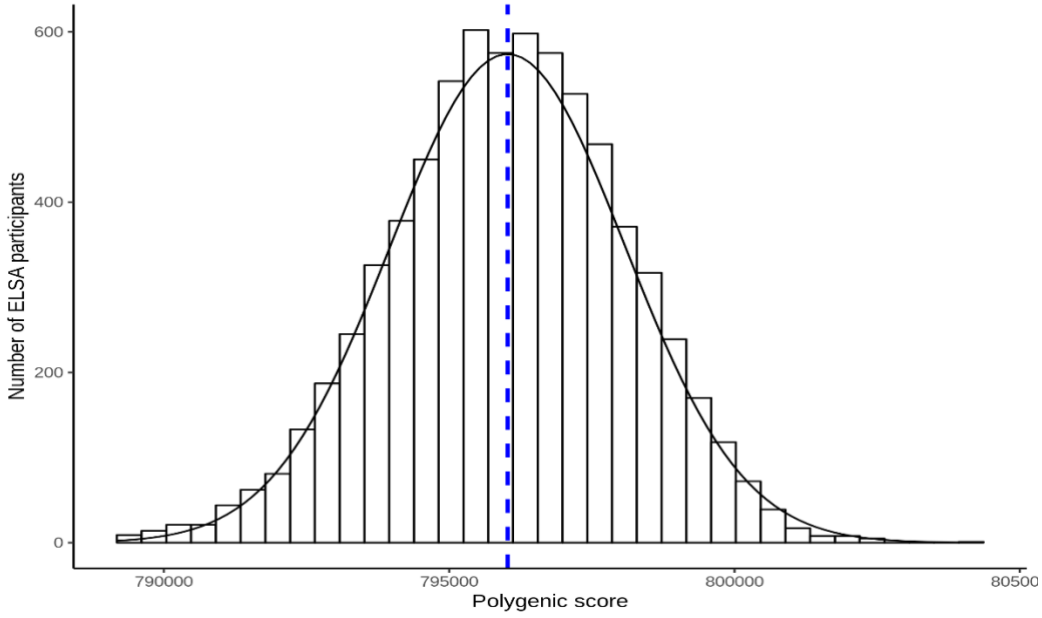
UKB, UK Biobank; CHARGE, Heart and Aging Research in Genomic Epidemiology consortium; CORNET, CORtisol NETwork consortium

3.3.9. SLEEP-RELATED BEHAVIOURS AND OUTCOMES

3.3.9.1. Insomnia Complaints (2017)

PGS for the Insomnia complaints in ELSA was calculated using the GWAS results from the UK Biobank including ~73 million genetic variants in 152,249 individuals[43]. The first ~50,000 samples were genotyped on the UK BiLEVE custom array, and the remaining ~100,000 samples were genotyped on the UK Biobank Axiom array. After standard quality control of the SNPs and samples, which was performed by UK Biobank, the data set comprised 641,018 autosomal SNPs in 113,006 samples of European ancestry for phasing and imputation. Imputation was performed with a reference panel that included the UK10K haplotype panel and the 1000 Genomes Project Phase 3 reference panel. Association tests were performed in SNPTEST using logistic regression with the covariates age, sex (for the full sample), genotyping array, the top five genetically determined PCs and additional PCs out of ten further ones that were associated with the phenotype (tested by logistic regression). The distribution of PGS for Insomnia Complaints in ELSA is depicted in **Figure 33**. The PGS contain 803,361 SNPs that overlapped between the ELSA genetic database and the GWAS meta-analysis; these SNPs were included in the PGS for Insomnia Complaints.

Figure 33. Distribution of PGS for Insomnia Complaints



3.3.9.2. Sleep duration (2017)

PGSs for sleep duration in ELSA using the GWAS summary statistics perfumed using the data from the UK Biobank[69]. Sleep duration was a self-reported phenotype where study participants were asked, “About how many hours sleep do you get in every 24 hours? (please include naps),” with responses in hour increments. Participant DNA was genotyped on two arrays, UK BiLEVE and UKB Axiom, with >95% common content. Genotypes for 152,736 samples passed sample quality control (~99.9% of total samples). Before imputation, 806,466 SNPs passed quality control in at least one batch (>99% of the array content). Imputation of autosomal SNPs was performed to a merged reference panel comprising the Phase 3 1000 Genomes Project and UK10K panels. Genetic association analysis for autosomes was

performed in SNPTEST with the 'expected' method using an additive genetic model adjusted for age, sex, ten principal components and genotyping array. The distribution of PGS for Sleep Duration in ELSA is depicted in **Figure 30**. A total of 948331 SNPs overlapped with the ELSA genetic database with the GWAS summary statistics and were included in the PGS for this phenotype.

3.3.9.3. Sleep duration, short-sleep, and long-sleep (2019)

The GWAS summary statistics for these phenotypes came from the UK Biobank study, which is a prospective study of >500,000 people living in the United Kingdom[70]. All people in the National Health Service registry who were aged 40-69 years and living <25 miles from a study centre were invited to participate between 2006 and 2010. In total, 503,325 participants were recruited from over 9.2 million invitations. For the current analysis, 24533 individuals of non-white ethnicity (as defined in genotyping and quality control) were excluded to avoid confounding effects. Study participants (n ~ 500,000) self-reported sleep duration at baseline assessment. Participants were asked: About how many hours sleep do you get in every 24 h? (please include naps), with responses in hour increments. Sleep duration was treated as a continuous variable and also categorized as either short (6 h or less), normal (7 or 8 h), or long (9 h or more) sleep duration. Extreme responses of less than 3 h or more than 18 h were excluded, and 'Do not know' or 'Prefer not to answer' responses were set to missing. Participants who self-reported any sleep medication were excluded. The results highlighted identify 78 loci for self-reported habitual sleep duration ($p < 5 \times 10-8$; 43 loci at $p < 6 \times 10-9$). Separate GWAS for short (<7 h; n = 106,192 cases) and long (≥9 h; n = 34,184 cases) sleep relative to 7–8 h sleep duration (n = 305,742 controls) highlighted 27 and 8 loci, respectively, of which 13 were independent from the 78 sleep duration loci

3.3.9.4. Insomnia (2019)

A meta-analysis of the GWAS results of insomnia and morningness in the UKB and 23andMe cohorts was performed using fixed-effects meta-analysis METAL, using SNP P values weighted by sample size[71]. The prevalence of insomnia was 28.3% in the UKB version 2 sample, 30.5% in the 23andMe sample, and 29.9% in the combined sample, which is in keeping with previous estimates for people of advanced age in the UK. Older people dominate the UKB (mean age = 56.7, s.d. = 8.0) and 23andMe (two-thirds of the sample older than 45, one-third older than 60 years of age) samples. Meta-analysis identified 11990 GWS SNPs ($P < 5 \times 10^{-8}$), represented by 248 independent lead SNPs ($r^2 < 0.1$), located in 202 genomic risk loci. All lead SNPs showed concordant signs of effect in both samples (Supplementary Fig. 3b). We confirmed two (chr2:66,785,180 and chr5:135,393,752) out of six previously reported loci for insomnia. Polygenic score (PGS) prediction in three randomly selected hold-out samples (n = 3 × 3,000) estimated the current results to explain up to 2.6% of the variance in insomnia.

3.3.9.5. Morningness, ease of waking up, naps, daytime dozing, and snoring

Morningness. The genome-wide meta-analysis on morningness included 434,835 subjects and 11,597,492 SNPs[71]. The genetic correlation between the two samples included in the meta-analysis (UKB, N=345552 and 23andMe, N=89283) was estimated at 0.92 (SE=0.02). The individual GWAS's showed some inflation in genetic signal (λ=1.603 for UKB and λ=1.253

for 23andMe) and mean χ2 statistic (1.815 and 1.302, respectively). The LD Score regression (LDSC) intercept was 1.046; (SE=0.011) for UKB and 1.007 (SE=0.008) for 23andMe. The two cohorts were meta-analyzed in METAL. The quantile-quantile (Q-Q) plot of the meta-analysed results also showed moderate inflation in λ (1.749) and mean χ2 statistic (2.073). The LDSC SNP-based heritability (h^2_{SNP}) of morningness was 0.186 (SE=0.006). The morningness GWAS analysis identified 16805 GWS SNPs ($P<5\times10^{-8}$), represented by 274 independent lead SNPs, which were mapped to 207 independent genomic loci

Ease of getting up. The genome-wide analysis for ease of getting up included 385,949 subjects and 10862568 SNPs[71]. All subjects were derived from the UKB sample. The Q-Q plot of the genome-wide analysis showed some inflation (λ=1.446) and mean χ2 statistic (1.586). The LDSC intercept (1.041; SE=0.010) was consistent with inflation due to true polygenicity and large sample size. The LDSC SNP-based heritability (h2SNP) of ease of getting up was 0.071 (SE=0.003). The ease of getting up GWAS analysis identified 7248 GWS SNPs ($P<5\times10^{-8}$), represented by 70 independent lead SNPs, which were mapped to 62 independent genomic loci.

Daytime Napping. The genome-wide analysis on daytime napping included 386,577 subjects and 10858887 SNPs[71]. All subjects were derived from the UKB sample. The QQ-plot of the genome-wide analysis showed some inflation (λ=1.159) and mean χ2 statistic (1.178). The LD score regression (LDSC) intercept (0.995; SE=0.007) was consistent with inflation due to true polygenicity and large sample size. The LDSC SNP-based heritability (h^2_{SNP}) of napping was 0.105 (SE=0.008). The daytime napping GWAS analysis identified 2,339 GWS SNPs ($P<5\times10^{-8}$), represented by 7 independent lead SNPs, which were mapped to 7 independent genomic loci.

Daytime Sleepiness/Dozing. The genome-wide analysis on dozing included 386,548 subjects and 10,820,725 SNPs[71]. All subjects were derived from the UKB sample. The Q-Q plot of the genome-wide analysis showed some inflation (λ=1.105) and mean χ2 statistic (1.107). The LDSC intercept (1.007; SE=0.007) was consistent with inflation due to true polygenicity and large sample size. The LDSC SNP-based heritability (h^2_{SNP}) of dozing was 0.091 (SE=0.010). The dozing GWAS analysis identified 9 GWS SNPs ($P<5\times10^{-8}$), represented by a single independent lead SNP (rs28600082, P=1.08×10-8) on chromosome 4, with surrounding SNPs located in non-coding RNA regions. The single risk locus for dozing was not shared with insomnia.

Snoring. The genome-wide analysis on snoring included 359,916 subjects and 10,862,568 SNPs[71]. All subjects were derived from the UKB sample. The QQ-plot of the genome-wide analysis showed some inflation (λ=1.358) and mean χ2 statistic (1.443). The LDSC intercept (1.010; SE=0.009) was consistent with inflation due to true polygenicity and large sample size. The LDSC SNP-based heritability (h^2_{SNP}) of snoring was 0.101 (SE=0.004). The snoring GWAS analysis identified 3,416 GWS SNPs ($P<5\times10^{-8}$), represented by 41 independent lead SNPs, which were mapped to 36 independent genomic loci.

Figure 34. Distribution of PGS for morningness, ease of waking up, naps, daytime dozing, and snoring

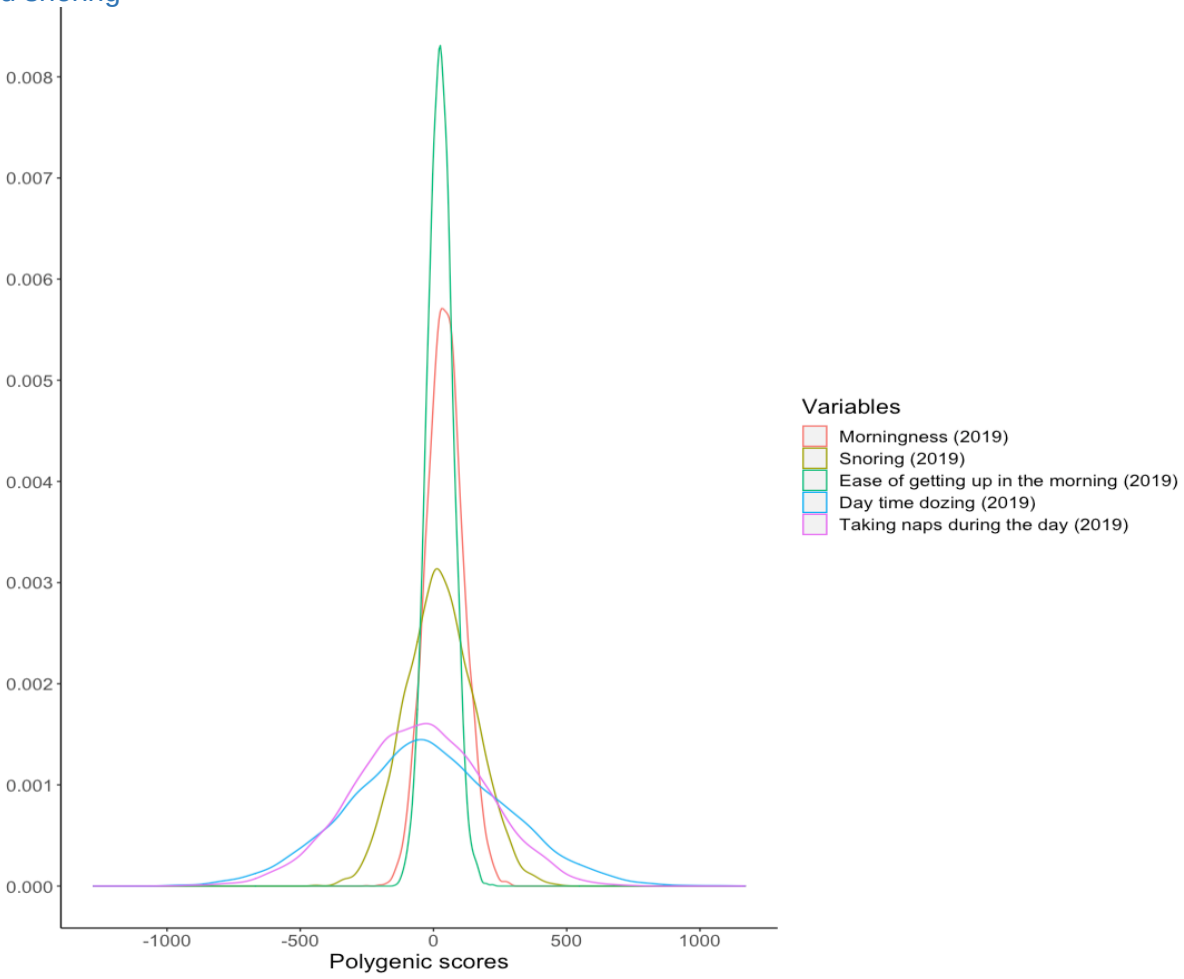


Table 18. The summary statistics for PGS for Sleep related traits

PGS	Sample Size	Minimum	Maximum	Range	Median	Mean	SE (mean)
Day time dozing (2019)	7183	-1275.9	1169.3	2445.2	-31.7	-25.4	3.45
Taking naps during the day (2019)	7183	-993.6	767.9	1761.6	-53.9	-55.0	2.86
Ease of getting up in the morning (2019)	7183	-140.6	223.4	364.1	24.3	24.3	0.57
Insomnia (2019)	7183	-308.7	542.0	850.7	105.1	106.3	1.40
Insomnia complaints (2017)	7183	789190	803945	14755	796054	796025.2	24.2
Morningness (2019)	7183	-251.9	286.1	537.9	41.5	42.1	0.81
Sleep duration (2017)	7183	5552.6	5665.2	112.6	5610.3	5610.2	0.18
Sleep duration (2019)	7183	-307.8	185.9	493.7	-41.1	-42.1	0.79
Sleep long (2019)	7183	7525.0	7674.2	149.3	7607.1	7607.1	0.20
Sleep short (2019)	7183	10407.5	10580.5	173.0	10490.2	10490.3	0.29
Snoring (2019)	7183	-446.9	560.0	1006.9	22.5	24.5	1.53

PGS, polygenic score; SE, standard error

Figure 35. Correlations between each sleep-related trait

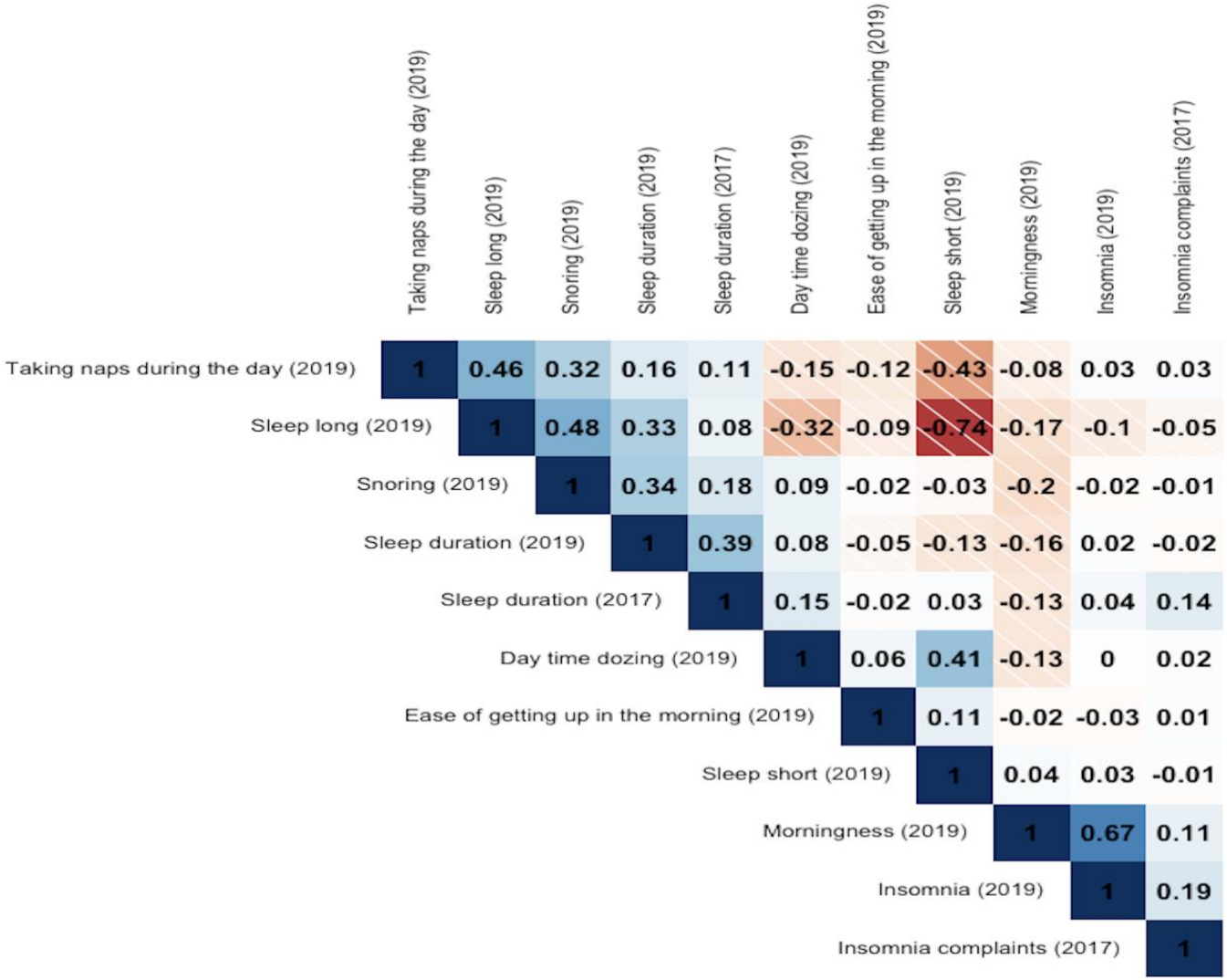


Table 19. Sources of the GWAS summary statistics used for these sleep-related traits

Phenotype	Consortium	GWAS SNPs	Overlapping with ELSA	GWAS meta-analysis citation	Source of base data
Insomnia complaints (2017)	UK Biobank	-	-	Hammerschlag et al. (2017)[43]	http://ctg.cncr.nl/software/summary_statistics.
Sleep duration (2017)	UK Biobank	-	-	Lane et al. (2017)[69]	Summary GWAS statistics will be made available at the UK Biobank website (http://biobank.ctsu.ox.ac.uk/).
Insomnia (2019)	UK Biobank	10,857,968	2,092,574	Jansen et al (2019)[71]	https://ctg.cncr.nl/software/summary_statistics
Sleep duration (2019)	UK Biobank & 23andMe	10,862,567	2,092,574	Dashti et al. (2019) [70]	Summary GWAS statistics are publicly available at the Sleep Disorder Knowledge Portal (http://sleepdisordergenetics.org/) and the UK Biobank website (http://biobank.ctsu.ox.ac.uk/).
Sleep long (2019)		14,661,601	6,246,221		
Sleep short (2019)		14,661,601	6,227,565		
Day time dozing (2019)		10,854,289	2,092,574	Jansen et al (2019)[71]	https://ctg.cncr.nl/software/summary_statistics
Taking naps during the day (2019)		10,820,724	2,092,574		
Ease of getting up in the morning (2019)		10,657,969	2,092,574		
Morningness (2019)		10,862,567	2,092,574		
Snoring (2019)		10,857,969	2,092,574		

3.3.10. REPRODUCTIVE FACTORS

3.3.10.1 Age at Menarche

PGSs for age at menarche were created using results from a 2014 study conducted by the Reproductive Genetics (ReproGen) consortium[72]. The ReproGen meta-analysis included 182,416 women of European descent from 57 studies imputed to HapMap Phase 2 CEU build 35 or 36 with a total of 2,441,815 autosomal SNPs. Birth year was the only covariate included to allow for the secular trends in menarche timing. The study reported 3,915 genome-wide significant SNPs. Of these, the authors identified 123 independent signals for age at menarche, which they assessed further in an independent sample of 8,689 women from the EPIC-InterAct study. The ReproGen GWAS summary statistics for this phenotype contained 2,441,815 SNPs; of these, 793,272 SNPs overlapped with the ELSA genetic database and were included in the PGS for Age at Menarche phenotype.

3.3.10.2. Age at Menopause

PGSs for age at menarche were created using results from a 2014 study conducted by the Reproductive Genetics (ReproGen) consortium[72]. The ReproGen meta-analysis included 182,416 women of European descent from 57 studies imputed to HapMap Phase 2 CEU build 35 or 36 with at total of 2,441,815 autosomal SNPs. Birth year was the only covariate included to allow for the secular trends in menarche timing. The study reported 3,915 genome-wide significant SNPs. Of these, the authors identified 123 independent signals for age at menarche, which they assessed further in an independent sample of 8,689 women from the EPIC-InterAct study. The PGSs contain 777,339 SNPs that overlapped between the ELSA genetic database and the GWAS meta-analysis; these SNPs were included in the PGS for this phenotype.

3.3.10.3. Age at first birth – Female & Male

PGSs for the Age at First Birth (AFB) for women and men were created using the GWAS summary statistics conducted by Barban et al. (2016)[73]. The total sample size of the meta-analysis for AFB was *n*=251,151. Cohorts uploaded results imputed using the HapMap 2 CEU (r22.b36) or 1000G reference sample. The analyses were adjusted for sex, birth year, and cohort specific covariates. The PGS for AFB for female participants contain 789,658 SNPs that overlapped between the ELSA genetic database and the GWAS meta-analysis; for the male participants, the PGS contained 787,685 SNPs that overlapped between the ELSA genetic database and the GWAS meta-analysis. These SNPs were included in the PGS for AFB phenotype.

3.3.10.4. Number of children ever born (NEB) – Female & Male

PGSs for the number of children ever born (NEB) for women and men were created using the GWAS summary statistics conducted by Barban et al. (2016)[73]. The total sample size of the meta-analysis was *n*=343,072 for NEB pooled. Cohorts uploaded

results imputed using the HapMap 2 CEU (r22.b36) or 1000G reference sample. The analyses were adjusted for sex, birth year, and cohort specific covariates. The PGS for NEB for female participants contains 793,718 SNPs that overlapped between the ELSA genetic database and the GWAS meta-analysis; for the male participants, the PGS contained 793,205 SNPs that overlapped between the ELSA genetic database and the GWAS meta-analysis. These SNPs were included in the PGS for NEB phenotype.

Figure 36. Distribution of PGS for reproductive factors

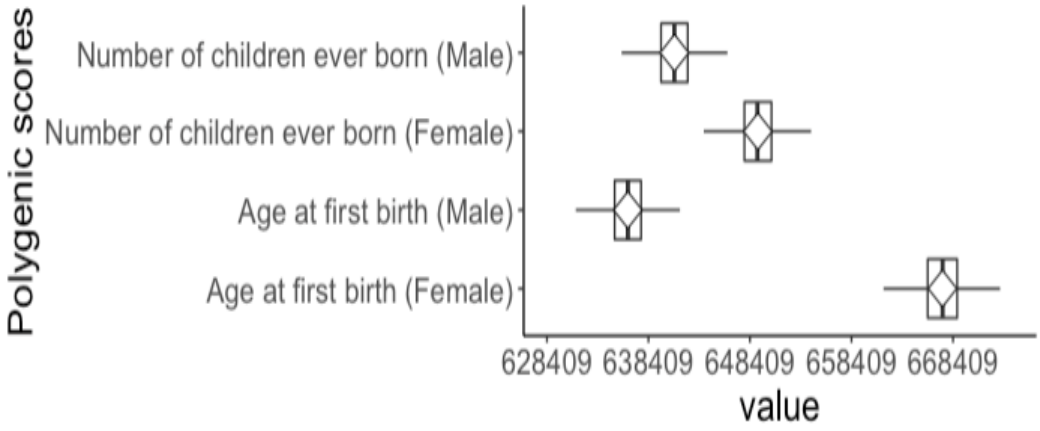


Table 20. The summary statistics for PGS reproductive factors

PGS	Gender	Sample size	Minimum	Maximum	Range	Median	Mean	SE (mean)
Age at Menopause	Female	3878	19882.8	20418.7	535.9	20179.8	20178.0	1.10
Age at Menarche	Male	3878	6123.6	6273.2	149.6	6192.6	6192.8	0.30
Age at first birth	Female	3878	659778	674336	14558	667389	667371.7	34.31
	Male	3305	628409	644380	15971	636385	636376.4	34.53
Number of children ever born	Female	3878	640948	655822	14874	649140	649185.0	31.96
	Male	3305	634040	647545	13505	640960	640967.4	34.10

PGS, polygenic score; SE, standard error

Table 21. Sources of the GWAS summary statistics used for PGS reproductive factors

Phenotype	Consortium	GWAS SNPs	Overlapping with ELSA	GWAS meta-analysis citation	Source of base data
Age at Menarche	ReproGen	2,441,815	793,272	Perry et al. (2014)[72]	http://www.reprogen.org/data_download.html (Menarche_Nature2014_GWASMetaResults_17122014.txt).
Age at Menopause	ReproGen	2,418,695	777,339		http://www.reprogen.org/data_download.html.
Age at first birth – Female	-	2,470,136	789,658	Barban et al. (2016)[73]	ftp://ftp.ebi.ac.uk/pub/databases/gwas/summary_statistics/BarbanN_27798627_GCST006045
Age at first birth – Male	-	2,465,140	787,685		ftp://ftp.ebi.ac.uk/pub/databases/gwas/summary_statistics/BarbanN_27798627_GCST006045
Number of children – Female	-	2,471,862	793,718		ftp://ftp.ebi.ac.uk/pub/databases/gwas/summary_statistics/BarbanN_27798627_GCST006047/
Number of children – Male	-	2,470,443	793,205		https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5695684/

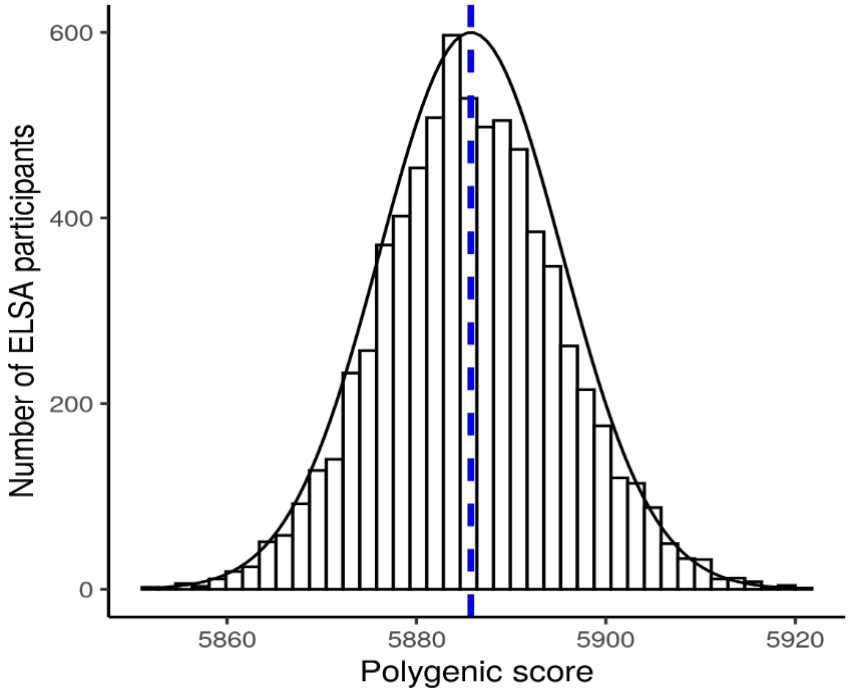
ReproGe; Reproductive Genetics consortium

3.3.11. INTELLIGENCE

3.3.11.1. Intelligence (2018)

PGSs for subjective intelligence were created using results from the largest genetic association study of intelligence to date that encompassed data from 14 independent epidemiological cohorts of European ancestry and 9,295,118 genetic variants passing quality control[74]. Intelligence was assessed using various neurocognitive tests, primarily gauging fluid domains of cognitive functioning. In this meta-analysis, 12,110 variants indexed by 242 lead SNPs in approximate linkage equilibrium (P_{GWAS} $<5\times10^{-8}$). The distribution of PGS for intelligence in ELSA is depicted in **Figure 37**. GWAS summary statistics contained 9,295,118 SNPs; of these, 1,269,550 SNPs overlapped with the ELSA genetic database and were included in the PGS for Intelligence phenotype.

Figure 37. Distribution of PGS for Intelligence (2018)

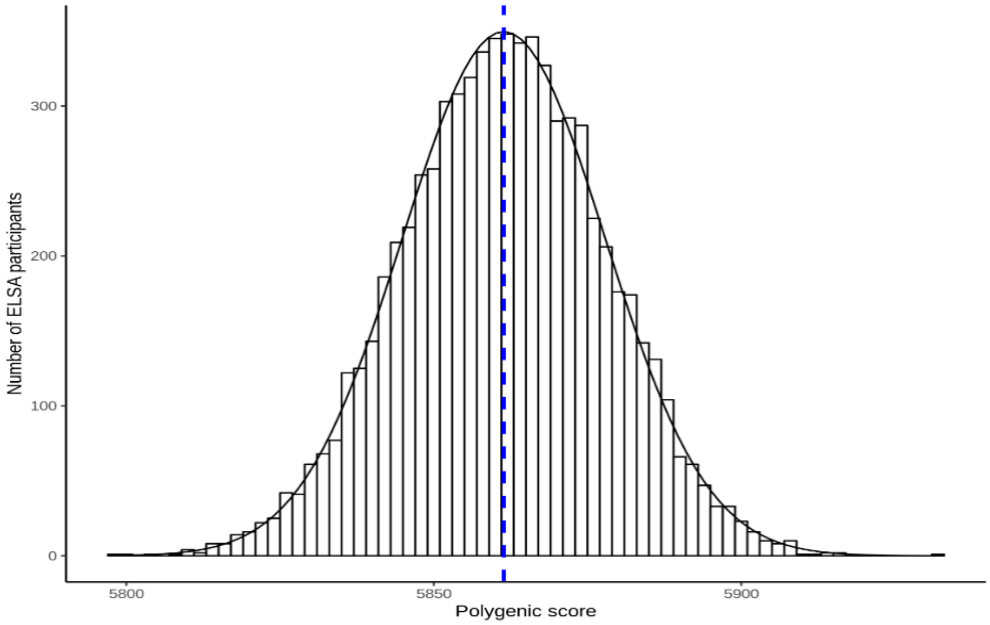


3.3.5.5. General Cognition (2015)

The PGSs for general cognition were created using results from a 2015 GWAS conducted across 31 cohorts by the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) consortium[75]. A total of 53,949 participants undertook multiple, diverse cognitive tests from which a general cognitive function phenotype was created within each cohort by principal component analysis. Thirteen genome-wide significant SNPs in three separate regions previously associated with neuropsychiatric phenotypes were reported. Adjustments for age, sex, and population stratification were included in study specific GWAS association analyses. Cohort-specific covariates - for example, site or familial relationships - were also fitted as required. The distribution of PGS for General Cognition in ELSA is depicted in **Figure**

38. A total of 2473946 SNPs were included in the CHARGE meta-analysis summary statistics. Of these, 795,327 SNPs overlapped with the ELSA genetic database and were included in the PGS for the general cognition phenotype.

Figure 38. Distribution of PGS for General Cognition (2015)



3.3.5.6. General Cognition (2018)

The PGSs for general cognition were created using results from a 2018 GWAS; it included 300,486 individuals of European ancestry from 57 population-based cohorts brought together by the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE), the Cognitive Genomics Consortium (COGENT) consortia, and UK Biobank[76]. All individuals were aged between 16 and 102 years. Exclusion criteria included clinical stroke (including self-reported stroke) or prevalent dementia. However, the ELSA study was part of the original analyses. We therefore requested that the consortia to repeat the analysis with ELSA removed. The PGSs for general cognition presented here are based on summary statistics that did not include ELSA. Genotype–phenotype association analyses were performed within each cohort, using an additive model, on imputed SNP dosage scores. Adjustments for age, sex, and population stratification were included in the model for each cohort. Cohort-specific covariates - for example, site or familial relationships - were also fitted as required. The distribution of PGS for General Cognition in ELSA is depicted in **Figure 39**. A total SNPs included in the meta-analysis summary statistics, 1348174 SNPs overlapped with the ELSA genetic database and were included in the PGS for the general cognition phenotype.

Figure 39. Distribution of PGS for General Cognition (2018)

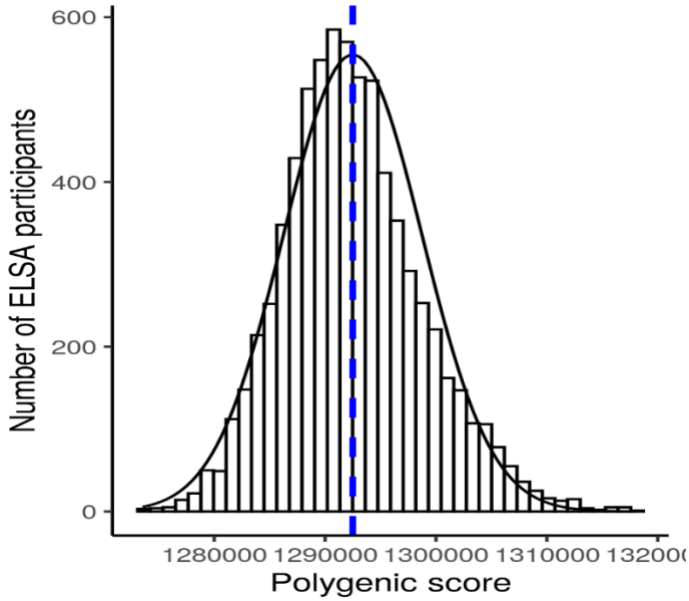


Table 22. Descriptive statistics for PGS for intelligence and general cognition

PGS	Minimum	Maximum	Range	Median	Mean	SE (mean)
Intelligence	5851.6	5920.5	68.9	5885.5	5885.7	0.11
General Cognition (2015)	5798.4	5932.8	134.3	5861.6	5861.4	0.19
General Cognition (2018)	1273530	1318000	44470	1291970.0	1292488.7	73.4

PGS, polygenic score; SE, standard error

Table 23. Sources of the GWAS summary statistics used for PGS intelligence and general cognition

Phenotype	Consortium	GWAS SNPs	Overlapping with ELSA	GWAS meta-analysis citation	Source of base data
Intelligence	UK Biobank, COGENT, RS,GENR,STR,S4S,HiQ/HRS,TEDS, DTR, IMAGEN, BLTS,NESCOG,GfG,STA	9,285,776	1,269,550	Savage et al (2018) [74]	Summary statistics are available for download from https://ctg.cncr.nl/
General cognitive function (2015)	CHARGE	2,473,946	795,327	Davies et al. (2015)[75]	https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000930.v6.p1
General cognitive function (2018)	CHARGE & COGENT	12,240,776	1,348,174	Davies et al (2018)[76]	On request from the authors

UK Biobank (UKB), COGENT, Cognitive Genomics Consortium; RS; Rotterdam Study; GENR, Generation R Study; STR, Swedish Twin Registry; S4S, Spit for Science; HiQ/HRS, High-IQ/Health and Retirement Study; TEDS, Twins Early Development Study; DTR Danish Twin Registry; IMAGEN; BLTS, Brisbane Longitudinal Twin Study; NESCOG Netherlands Study of Cognition, Environment, and Genes; GfG, Genes for Good; STSA, Swedish Twin Studies of Aging; CHARGE, Heart and Aging Research in Genomic Epidemiology consortium; CORNET, CORtisol NETwork consortium

3.3.12. LONGEVITY

3.3.12.1. Longevity (2015)

The longevity PGSs were created using summary statistics from a 2015 GWAS conducted by the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) consortia[77]. The GWAS summary statistics for this phenotype were obtained from the GRASP (Genome-Wide Repository of Associations Between SNPs and Phenotypes) [78]. The GWAS meta-analysis on longevity used the sample of 6,036 longevity cases and 3,757 controls accumulated across 11 studies. The data was imputed to ~2.5 million SNPs using the HapMap 22 CEU (Build 36) genotyped samples as a reference. Logistic regression was used to test each SNP for association with longevity using an additive model adjusting for sex and PCs to adjust for population stratification. None of the SNP-longevity associations reached the genome-wide significance threshold of 5×10^{-8} in the discovery phase. Suggestive evidence was found for the involvement of SNPs near CADM2 and GRIK2, and the associations of APOE and FOXO3 with longevity were confirmed. A total of 2588525 SNPs were included in the summary statistics. Of these, 757472 SNPs overlapped with the ELSA genetic database and were included in the PGS for this phenotype.

3.3.12.2. Longevity (2019)

In this GWAS meta-analysis[79], cohorts that participated in one or more of the previously published GWA studies on longevity were included; these were:[77, 80, 81]. Cases were individuals who lived to an age above the 90th percentile based on cohort life tables from census data from the appropriate country, sex, and birth cohort. Controls were individuals who died at or before the age at the 60th percentile or whose age at the last follow-up visit was at or before the 60th percentile age. Hence, the number of selected cases and controls is defined by the ages of their birth cohort corresponding to the 60th or 90th percentile age and is independent of the study population used (i.e., the number of controls and cases within a study population is not based on the percentiles of that specific population, but instead on that of their birth cohorts). Here, the study encompassed 11,262 cases surviving at or beyond the age corresponding to the 90th survival percentile, respectively, and 25,483 controls whose age at death or at last contact was at or below the age corresponding to the 60th survival percentile.

3.3.12.3. Parental death (2017)

Genetic data was available on 488,377 UK Biobank participants after genotype calling and quality control performed centrally by the UK Biobank team. In total, 451,447 participants identified as ‘white European’ through self-report and verified through principal components analysis based on genotypes were selected[82]. Briefly, principal components were generated in the 1000 Genomes Cohort using high-confidence SNPs to obtain their individual loadings. These loadings were then used to project all the UK Biobank samples into the same principal component space and individuals were then clustered using principal components 1 to 4. Related individuals were identified through kinship analysis, although these participants were included in genome-wide analysis using BOLT-LMM, those related to the third-degree or closer were excluded in sensitivity analyses. Imputation of 39,235,157 genetic variants from the Haplotype Reference Consortium panel was performed using IMPUTE centrally by the UK Biobank team. After filtering for variants with MAF ≥0.1%, missingness <1.5%, imputation

quality >0.1 and with Hardy-Weinberg equilibrium (HWE) $P>1\times10^{-6}$ within the white British participants 11,516,125 imputed autosomal variants were eligible for the analyses. Additionally, data were directly utilized from the microarrays for variants on the X (n=19,381) and Y (n=284) chromosomes, and on the mitochondrial genome (n=135), which were unavailable in the imputed dataset. Participants were asked the age at which their parents had died (or their current age if still alive). Analyses were performed separately on mother's age at death and father's age at death, and also on a combined phenotype. To reduce the effect of higher ages at death of mothers (compared to the fathers) the mothers and fathers age at deaths were z-transformed before combining the z-scores into a single summed phenotype. Offspring of parents who died prematurely were excluded because the cause of death of the participant's parents was not asked, so we could not exclude accidental deaths explicitly.

Table 24. The descriptive statistics for PGS longevity

Phenotypes	Sample Size	Minimum	Maximum	Range	Median	Mean	SE (mean)
Longevity (2015)	7183	619.4	1334.4	714.9	976.5	977.2	1.03
Longevity (2019)	7183	36405.9	36951.1	545.2	36700.9	36699.0	0.92
Parental death (2017)	7183	7689.7	7814.4	124.7	7753.6	7753.4	0.21

PGS, polygenic score; SE, standard error

Table 25. Sources of the GWAS summary statistics used for PGS longevity

Phenotype	Consortium	GWAS SNPs	Overlapping with ELSA	GWAS meta-analysis citation	Source of base data
Longevity (2015)	-	-	-	Broer et al (2015)[66]	https://grasp.nhlbi.nih.gov/FullResults.aspx
Longevity (2019)	-	7,508,009	1,329,210	Deelen et al. (2019)[79]	https://grasp.nhlbi.nih.gov/FullResults.aspx; https://www.longevitygenomics.org/downloads
Parental death (2017)	UK Biobank	9,085,648	1,329,730	Pilling et al (2017)[82]	https://doi.org/10.6084/m9.figshare.5439382.v1

4. SET UP

4.1. Download the PGSs in ELSA

By downloading these freely provided datasets, you agree to use their contents only for research and statistical purposes, making no effort to identify the respondents. The generated PGSs are available for download in three data formats (STATA, SPSS, and EXCEL):

1. List_PGS_SCORES_ELSA_JULY_2022.dta
2. List_PGS_SCORES_ELSA_JULY_2022.dav
3. List_PGS_SCORES_ELSA_JULY_2022.xlsx

All data files are keyed on unique identifier (IDAUNIQ).

4.2. Why to use principal component in association analyses?

Population stratification occurs when the differences in the allele frequency between cases and controls are due to systematic ancestry differences leading to spurious associations in studies[7]. To account for any ancestry differences in genetic structures that could bias the results, it is advisable to adjust the association analyses for principal components (PCs) (for more detail, please refer to page 13). Some studies adjust for all 10 PCs; others tend to use the first 4 PCs; while some recommend checking whether ancestry PCs associate with the phenotypes under investigation. If they do, or the cohort under investigation has known issues with stratification, then it is advisable to adjust for these PCs. Ultimately, the researchers will need to make the decision whether to use PCs in their analyses, and if so, how many.

In ELSA we have generated 10 ancestry principal components. These are provided in three data formats data formats (STATA, SPSS, and EXCEL):

1. Principal_Components ELSA_JULY_2022.dta
2. Principal_Components ELSA_JULY_2022.sav
3. Principal_Components ELSA_ ULY_2022.xlsx

All data files are keyed on unique identifier (IDAUNIQ).

4.3.Additional data available

To improve genome coverage, we imputed untyped quality-controlled using the University of Michigan Imputation Server[83]. To estimate genotypes that were not assayed, imputation was performed on the Michigan Imputation Server[83] running SHAPEIT for pre-phasing[84], and Minimac3 for imputation[85, 86] to the Haplotype Reference Consortium (HRC.r1-1.GRCh37)[83, 87]; all variants align to human genome build 19 (hg19). Post-imputation, we kept variants that were genotyped or imputed at INFO>0.95, in low linkage disequilibrium (R^2<0.1) and with Hardy-Weinberg Equilibrium p-value>10^{-5}. After the sample quality control 7179780 variants were retained for further analyses. To account for any ancestry differences in genetic structures that could bias results, principal components analysis was conducted retaining top principal components (PCs)[7]. After the sample quality control 7,179,780 variants were retained for further analyses. These data are available for researchers if they wish to use them in their work.

4.4. If You Need to Know More

This document is intended to serve as a brief overview to provide guidelines for using the *ELSA Polygenic Scores* data product in the ELSA study. If you have questions or concerns

that are not adequately covered here, or if you have any comments, please contact us. We will do our best to provide answers.

4.5. Contact Information

If you need to contact us, you may do so by one of the methods listed below.

Email: Please send your concerns or requests for further information to Dr Olesya Ajnakina using the email address: o.ajnakina@ucl.ac.uk

Post: Please send your concerns or requests for further information to Dr Olesya Ajnakina using the postal address:

Department of Behavioural Science and Health
Institute of Epidemiology and Health Care
University College London
Postal address: UCL, Gower Street, London WC1E 6BT

5. REFERENCES

1. Ware EB, et al., *Method of Construction Affects Polygenic Score Prediction of Common Human Trait.* BiorXiv, 2017: p. 1-13.
2. Wray, N.R., et al., *Research review: Polygenic methods and their application to psychiatric traits.* J Child Psychol Psychiatry, 2014. **55**(10): p. 1068-87.
3. Purcell, S.M., et al., *Common polygenic variation contributes to risk of schizophrenia and bipolar disorder.* Nature, 2009. **460**(7256): p. 748-52.
4. Dudbridge, F., *Power and predictive accuracy of polygenic risk scores.* PLoS Genet, 2013. **9**(3): p. e1003348.
5. Hardy, J. and A. Singleton, *Genomewide association studies and human disease.* N Engl J Med, 2009. **360**(17): p. 1759-68.
6. So, H.C. and P.C. Sham, *Improving polygenic risk prediction from summary statistics by an empirical Bayes approach.* Sci Rep, 2017. **7**: p. 41262.
7. Price, A.L., et al., *Principal components analysis corrects for stratification in genome-wide association studies.* Nat Genet, 2006. **38**(8): p. 904-9.
8. Mavaddat, N., et al., *Prediction of breast cancer risk based on profiling with common genetic variants.* J Natl Cancer Inst, 2015. **107**(5).
9. Yang, J., et al., *GCTA: a tool for genome-wide complex trait analysis.* Am J Hum Genet, 2011. **88**(1): p. 76-82.
10. Mullins, N., et al., *Polygenic interactions with environmental adversity in the aetiology of major depressive disorder.* Psychol Med, 2016. **46**(4): p. 759-70.
11. Natarajan, P., et al., *Polygenic Risk Score Identifies Subgroup With Higher Burden of Atherosclerosis and Greater Relative Benefit From Statin Therapy in the Primary Prevention Setting.* Circulation, 2017. **135**(22): p. 2091-2101.
12. Steptoe, A., et al., *Cohort profile: the English longitudinal study of ageing.* Int J Epidemiol, 2013. **42**(6): p. 1640-8.
13. Sonnega, A., et al., *Cohort Profile: the Health and Retirement Study (HRS).* Int J Epidemiol, 2014. **43**(2): p. 576-85.
14. Marees, A.T., et al., *A tutorial on conducting genome-wide association studies: Quality control and statistical analysis.* Int J Methods Psychiatr Res, 2018. **27**(2): p. e1608.
15. Chang, C.C., et al., *Second-generation PLINK: rising to the challenge of larger and richer datasets.* Gigascience, 2015. **4**: p. 7.
16. Huff, C.D., et al., *Maximum-likelihood estimation of recent shared ancestry (ERSA).* Genome Res, 2011. **21**(5): p. 768-74.
17. Laurie, C.C., et al., *Quality control and quality assurance in genotypic data for genome-wide association studies.* Genet Epidemiol, 2010. **34**(6): p. 591-602.
18. Anderson, C.A., et al., *Data quality control in genetic case-control association studies.* Nat Protoc, 2010. **5**(9): p. 1564-73.
19. Novembre, J., et al., *Genes mirror geography within Europe.* Nature, 2008. **456**(7218): p. 98-101.
20. Wang, D., et al., *Comparison of methods for correcting population stratification in a genome-wide association study of rheumatoid arthritis: principal-component analysis versus multidimensional scaling.* BMC Proc, 2009. **3 Suppl 7**: p. S109.
21. Shing Wan Choi, T.S.H.M., Paul O'Reilly, *A guide to performing Polygenic Risk Score analyses.* bioRciv, 2018: p. 1-22.
22. Okbay, A., et al., *Genome-wide association study identifies 74 loci associated with educational attainment.* Nature, 2016. **533**(7604): p. 539-42.
23. Euesden, J., C.M. Lewis, and P.F. O'Reilly, *PRSice: Polygenic Risk Score software.* Bioinformatics, 2015. **31**(9): p. 1466-8.

24. van den Berg, S.M., et al., *Meta-analysis of Genome-Wide Association Studies for Extraversion: Findings from the Genetics of Personality Consortium.* Behav Genet, 2016. **46**(2): p. 170-82.
25. de Moor, M.H., et al., *Meta-analysis of genome-wide association studies for personality.* Mol Psychiatry, 2012. **17**(3): p. 337-49.
26. Okbay, A., et al., *Genetic variants associated with subjective well-being, depressive symptoms, and neuroticism identified through genome-wide analyses.* Nat Genet, 2016. **48**(6): p. 624-33.
27. al., T.G.P.C.e., *An integrated map of genetic variation from 1,092 human genomes. .* Nature 2012. **491**: p. 56–65.
28. Lee, J.J., et al., *Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals.* Nat Genet, 2018. **50**(8): p. 1112-1121.
29. Hill, W.D., et al., *Molecular Genetic Contributions to Social Deprivation and Household Income in UK Biobank.* Curr Biol, 2016. **26**(22): p. 3083-3089.
30. Lambert, J.C., et al., *Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease.* Nat Genet, 2013. **45**(12): p. 1452-8.
31. Jansen, I.E., et al., *Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk.* Nat Genet, 2019. **51**(3): p. 404-413.
32. Ripke, S., et al., *A mega-analysis of genome-wide association studies for major depressive disorder.* Mol Psychiatry, 2013. **18**(4): p. 497-511.
33. Sudlow, C., et al., *UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age.* PLoS Med, 2015. **12**(3): p. e1001779.
34. Wray, N.R., et al., *Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression.* Nat Genet, 2018. **50**(5): p. 668-681.
35. Otowa, T., et al., *Meta-analysis of genome-wide association studies of anxiety disorders.* Mol Psychiatry, 2016. **21**(10): p. 1485.
36. *Biological insights from 108 schizophrenia-associated genetic loci.* Nature, 2014. **511**(7510): p. 421-7.
37. Stahl, E.A., et al., *Genome-wide association study identifies 30 loci associated with bipolar disorder.* Nat Genet, 2019. **51**(5): p. 793-803.
38. 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, 2021. **53**(6): p. 817-829.
39. Demontis, D., et al., *Discovery of the first genome-wide significant risk loci for attention deficit/hyperactivity disorder.* Nat Genet, 2019. **51**(1): p. 63-75.
40. Gao, J., et al., *Genome-Wide Association Study of Loneliness Demonstrates a Role for Common Variation.* Neuropsychopharmacology, 2017. **42**(4): p. 811-821.
41. Day, F.R., K.K. Ong, and J.R.B. Perry, *Elucidating the genetic basis of social interaction and isolation.* Nat Commun, 2018. **9**(1): p. 2457.
42. *Meta-analysis of GWAS of over 16,000 individuals with autism spectrum disorder highlights a novel locus at 10q24.32 and a significant overlap with schizophrenia.* Mol Autism, 2017. **8**: p. 21.
43. Hammerschlag, A.R., et al., *Genome-wide association analysis of insomnia complaints identifies risk genes and genetic overlap with psychiatric and metabolic traits.* Nat Genet, 2017. **49**(11): p. 1584-1592.
44. Pappa, I., et al., *A genome-wide approach to children's aggressive behavior: The EAGLE consortium.* Am J Med Genet B Neuropsychiatr Genet, 2016. **171**(5): p. 562-72.
45. Benke, K.S., et al., *A genome-wide association meta-analysis of preschool internalizing problems.* J Am Acad Child Adolesc Psychiatry, 2014. **53**(6): p. 667-676.e7.
46. Dalvie, S., et al., *Genomic influences on self-reported childhood maltreatment.* Transl Psychiatry, 2020. **10**(1): p. 38.
47. Schunkert, H., et al., *Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease.* Nat Genet, 2011. **43**(4): p. 333-8.

48. Fall, T., et al., *Genome-wide association study of coronary artery disease among individuals with diabetes: the UK Biobank.* Diabetologia, 2018. **61**(10): p. 2174-2179.
49. Morris, A.P., et al., *Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes.* Nat Genet, 2012. **44**(9): p. 981-90.
50. Xue, A., et al., *Genome-wide association analyses identify 143 risk variants and putative regulatory mechanisms for type 2 diabetes.* Nat Commun, 2018. **9**(1): p. 2941.
51. Okada, Y., et al., *Genetics of rheumatoid arthritis contributes to biology and drug discovery.* Nature, 2014. **506**(7488): p. 376-81.
52. Consortium, C.D., *A comprehensive 1000 Genomes–based genome-wide association meta-analysis of coronary artery disease.* Nature, 2015. **47**: p. 1121–1130.
53. Gormley, P., et al., *Meta-analysis of 375,000 individuals identifies 38 susceptibility loci for migraine.* Nat Genet, 2016. **48**(8): p. 856-66.
54. Johnston, K.J.A., et al., *Genome-wide association study of multisite chronic pain in UK Biobank.* PLoS Genet, 2019. **15**(6): p. e1008164.
55. Tikkanen, E., et al., *Biological Insights Into Muscular Strength: Genetic Findings in the UK Biobank.* Sci Rep, 2018. **8**(1): p. 6451.
56. Ben-Avraham, D., et al., *The complex genetics of gait speed: genome-wide meta-analysis approach.* Aging (Albany NY), 2017. **9**(1): p. 209-246.
57. Wood, A.R., et al., *Defining the role of common variation in the genomic and biological architecture of adult human height.* Nat Genet, 2014. **46**(11): p. 1173-86.
58. Locke, A.E., et al., *Genetic studies of body mass index yield new insights for obesity biology.* Nature, 2015. **518**(7538): p. 197-206.
59. Yengo, L., et al., *Meta-analysis of genome-wide association studies for height and body mass index in approximately 700000 individuals of European ancestry.* Hum Mol Genet, 2018. **27**(20): p. 3641-3649.
60. Shungin, D., et al., *New genetic loci link adipose and insulin biology to body fat distribution.* Nature, 2015. **518**(7538): p. 187-196.
61. *Genome-wide meta-analyses identify multiple loci associated with smoking behavior.* Nat Genet, 2010. **42**(5): p. 441-7.
62. 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.* Nat Genet, 2019. **51**(2): p. 237-244.
63. Schumann, G., et al., *KLB is associated with alcohol drinking, and its gene product beta-Klotho is necessary for FGF21 regulation of alcohol preference.* Proc Natl Acad Sci U S A, 2016. **113**(50): p. 14372-14377.
64. Bolton, J.L., et al., *Genome wide association identifies common variants at the SERPINA6/SERPINA1 locus influencing plasma cortisol and corticosteroid binding globulin.* PLoS Genet, 2014. **10**(7): p. e1004474.
65. Ligthart, S., et al., *Genome Analyses of >200,000 Individuals Identify 58 Loci for Chronic Inflammation and Highlight Pathways that Link Inflammation and Complex Disorders.* Am J Hum Genet, 2018. **103**(5): p. 691-706.
66. Said, S., et al., *Genetic analysis of over half a million people characterises C-reactive protein loci.* Nat Commun, 2022. **13**(1): p. 2198.
67. Vuckovic, D., et al., *The Polygenic and Monogenic Basis of Blood Traits and Diseases.* Cell, 2020. **182**(5): p. 1214-1231.e11.
68. Chen, M.H., et al., *Trans-ethnic and Ancestry-Specific Blood-Cell Genetics in 746,667 Individuals from 5 Global Populations.* Cell, 2020. **182**(5): p. 1198-1213.e14.
69. Lane, J.M., et al., *Genome-wide association analyses of sleep disturbance traits identify new loci and highlight shared genetics with neuropsychiatric and metabolic traits.* Nat Genet, 2017. **49**(2): p. 274-281.

70. Dashti, H.S., et al., *Genome-wide association study identifies genetic loci for self-reported habitual sleep duration supported by accelerometer-derived estimates.* Nat Commun, 2019. **10**(1): p. 1100.
71. Jansen, P.R., et al., *Genome-wide analysis of insomnia in 1,331,010 individuals identifies new risk loci and functional pathways.* Nat Genet, 2019. **51**(3): p. 394-403.
72. Perry, J.R., et al., *Parent-of-origin-specific allelic associations among 106 genomic loci for age at menarche.* Nature, 2014. **514**(7520): p. 92-97.
73. Barban, N., et al., *Genome-wide analysis identifies 12 loci influencing human reproductive behavior.* Nat Genet, 2016. **48**(12): p. 1462-1472.
74. Savage, J.E., et al., *Genome-wide association meta-analysis in 269,867 individuals identifies new genetic and functional links to intelligence.* Nat Genet, 2018. **50**(7): p. 912-919.
75. Davies, G., et al., *Genetic contributions to variation in general cognitive function: a meta-analysis of genome-wide association studies in the CHARGE consortium (N=53949).* Mol Psychiatry, 2015. **20**(2): p. 183-92.
76. Davies, G., et al., *Study of 300,486 individuals identifies 148 independent genetic loci influencing general cognitive function.* Nat Commun, 2018. **9**(1): p. 2098.
77. Broer, L., et al., *GWAS of longevity in CHARGE consortium confirms APOE and FOXO3 candidacy.* J Gerontol A Biol Sci Med Sci, 2015. **70**(1): p. 110-8.
78. Leslie, R., C.J. O'Donnell, and A.D. Johnson, *GRASP: analysis of genotype-phenotype results from 1390 genome-wide association studies and corresponding open access database.* Bioinformatics, 2014. **30**(12): p. i185-94.
79. Deelen, J., et al., *A meta-analysis of genome-wide association studies identifies multiple longevity genes.* Nat Commun, 2019. **10**(1): p. 3669.
80. Deelen, J., et al., *Genome-wide association meta-analysis of human longevity identifies a novel locus conferring survival beyond 90 years of age.* Hum Mol Genet, 2014. **23**(16): p. 4420-32.
81. Zeng, Y., et al., *Novel loci and pathways significantly associated with longevity.* Sci Rep, 2016. **6**: p. 21243.
82. Pilling, L.C., et al., *Human longevity: 25 genetic loci associated in 389,166 UK biobank participants.* Aging (Albany NY), 2017. **9**(12): p. 2504-2520.
83. Das, S., et al., *Next-generation genotype imputation service and methods.* Nat Genet, 2016. **48**(10): p. 1284-1287.
84. Delaneau, O., J.F. Zagury, and J. Marchini, *Improved whole-chromosome phasing for disease and population genetic studies.* Nat Methods, 2013. **10**(1): p. 5-6.
85. Fuchsberger, C., G.R. Abecasis, and D.A. Hinds, *minimac2: faster genotype imputation.* Bioinformatics, 2015. **31**(5): p. 782-4.
86. Howie, B., et al., *Fast and accurate genotype imputation in genome-wide association studies through pre-phasing.* Nat Genet, 2012. **44**(8): p. 955-9.
87. McCarthy, S., et al., *A reference panel of 64,976 haplotypes for genotype imputation.* Nat Genet, 2016. **48**(10): p. 1279-83.

6. SUPPLEMENTARY MATERIAL

Supplementary Figure 1. Depicts distribution of 10 principal components once 65 individuals with ancestral admixture were removed from the sample.