DOCUMENTATION REPORT

THE ENGLISH LONGITUDINAL STUDY OF AGEING (ELSA)

POLYGENIC SCORES 2019

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 7HB

TABLE OF CONTENTS **PAGE**

LIST OF FIGURES	PAGE

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 30 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 2.

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 predictions 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.

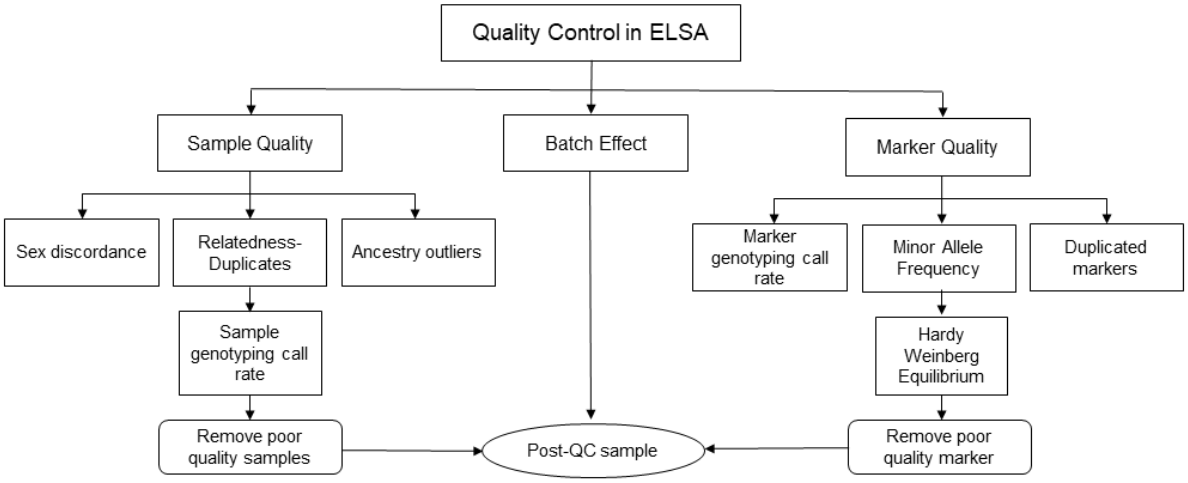
2.3. Genotyping Process

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 software [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. 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 2*).

2.4.4. 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. The summary of full QC procedure at each step is provided in *Supplementary Table 1*.

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 publically 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 publically 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. The details of the GWAS summary statistics used for the phenotypes described in this document can be found in *Supplementary Table 2*.

3.3. PGSs: 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 summary statistics from this GWAS meta-analysis are publicly available, the link to which can be found in Supplementary Table 2. The full 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 2.9×10^{-7} for a SNP located on chromosome 2. There were 74 SNPs with *p*-values $<1 \times 10^{-5}$.

The distribution of the PGS for Extraversion in the ELSA study is depicted in **Figure 2**; the summary statistics for the PGS for Extraversion are provided in **Table 1**. 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]. This GWAS was part of the Genetics of Personality Consortium (GPC) and is publicly available, the link to which can be found in Supplementary Table 2. This 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.

The distributions of the PGS for Agreeableness, Openness to Experience and Conscientiousness in the ELSA study are depicted in **Figure 2**; the summary statistics for the PGSs for these personality types are provided in **Table 1**.

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]. This GWAS was part of the Social Science Genetic Association Consortium (SSGAC) and is publicly available (Supplementary Table 2). 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.

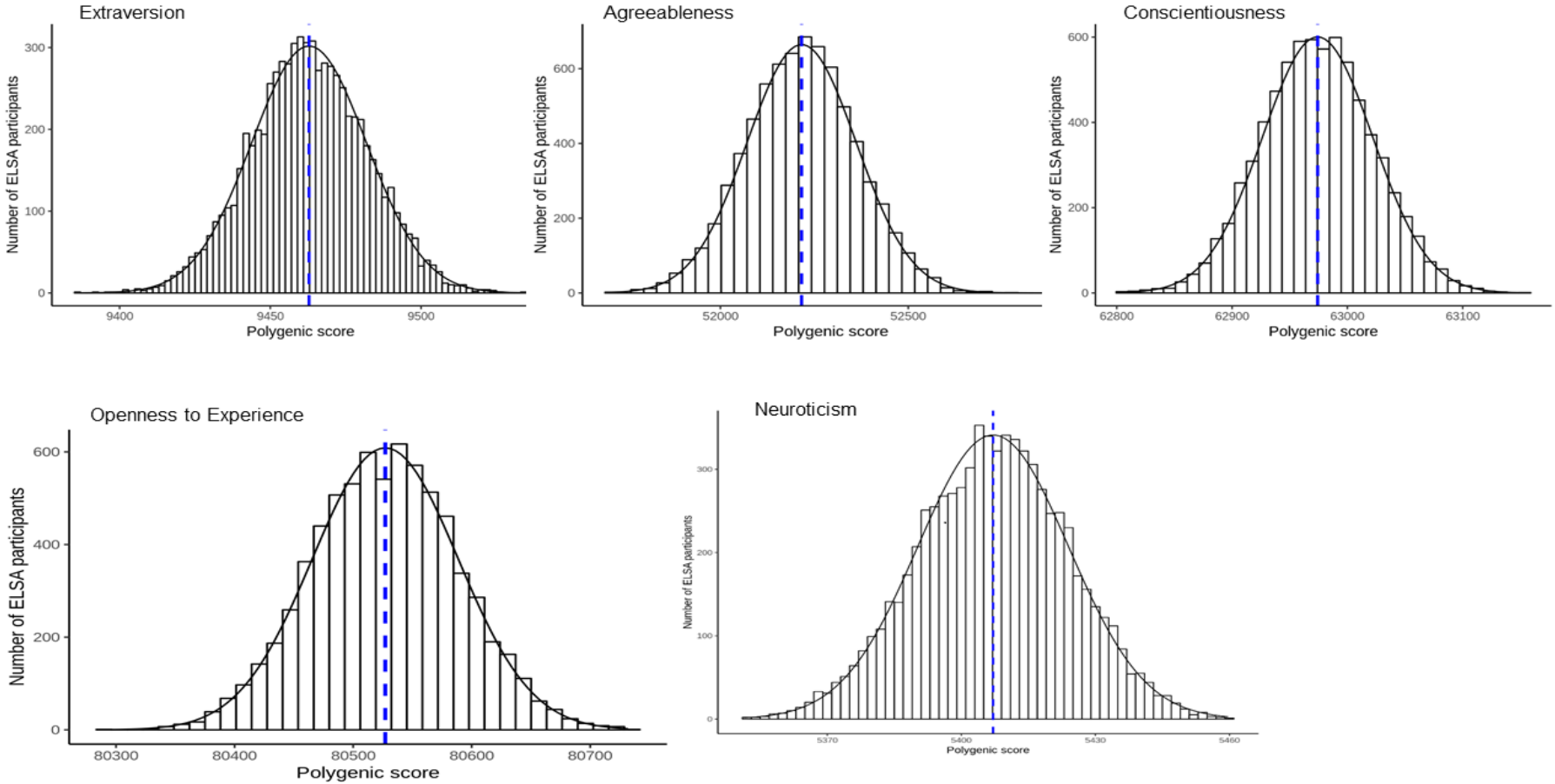
The distribution of the PGS for Neuroticism in the ELSA study is depicted in **Figure 2**; the summary statistics for the PGS for Neuroticism in ELSA are provided in **Table 1**. The GWAS for Neuroticism contained 6,524,432 SNPs; of these, 1,191,041 SNPs overlapped with the ELSA target data and were included in the PGSs for Neuroticism.

Table 1*. 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

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.

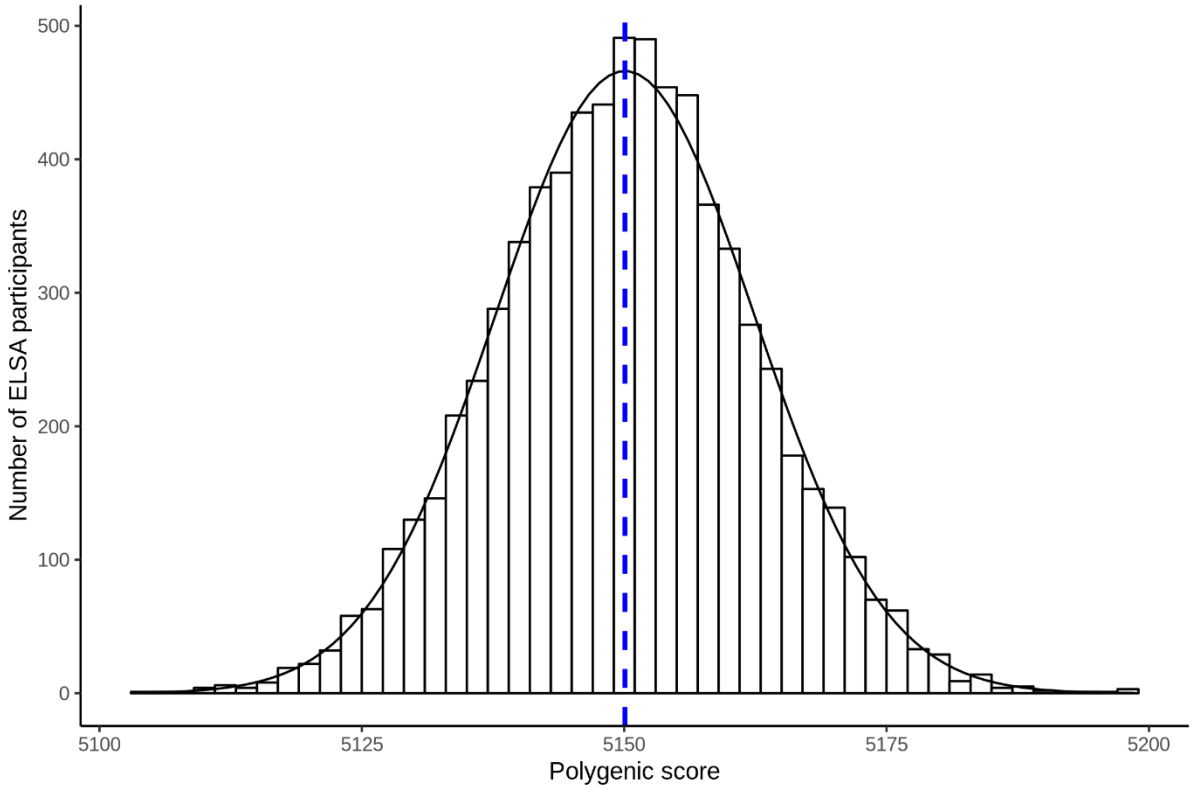
3.3.2.1.1. Educational Attainment - 2 (EA2)

PGSs for EA-2 were created using results from a 2016 study excluding 23andMe results (due to data use agreements)[22]. This GWAS was part of the Social Science Genetic Association Consortium (SSGAC) and is publicly available, the link to which can be found in Supplementary Table 2.

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. 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 summary statistics for PGS for EA-2 are provided in **Table 2**. 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 depict the mean

3.3.2.1.2. Educational Attainment - 3 (EA3)

Similarly 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]. The 2018 SSGAC GWAS meta-analysis files for EA are publicly available on their data download page (Supplementary Table 2). 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.

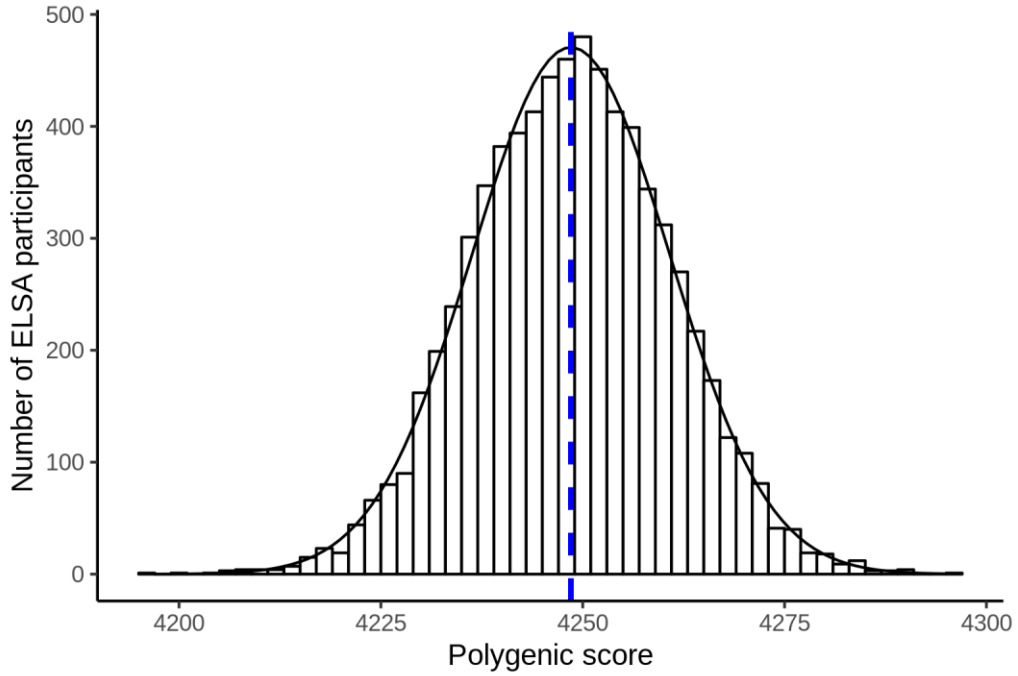
The distribution of PGS for EA-3 in the ELSA study is depicted in **Figure 4**; the summary statistics for PGS for EA-3 are provided in **Table 2**. The SSGAC GWAS 2018 contained 10,101,242 SNPs; of these, 1,325,851SNPs overlapped with the ELSA target data and were included in the PGSs for EA.

Table 2*. The summary statistics for PGS for EA-2 and EA-3*

PGSs	Sample Size	Minimum	Maximum	Range	Median	Mean	SE (mean)
EA-2	7183	5104.3	5198.5	94.2	5150.2	5150.1	0.15
EA-3	7183	4197.0	4296.7	99.7	4248.7	4248.5	0.14

PGS, polygenic score; SE, standard error

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. Full details can be found at http://biobank.ctsu.ox.ac.uk/crystal/refer.cgi?id=157020. 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 summary statistics from this GWAS meta-analysis are publicly available, the link to which can be found in Supplementary Table 2.

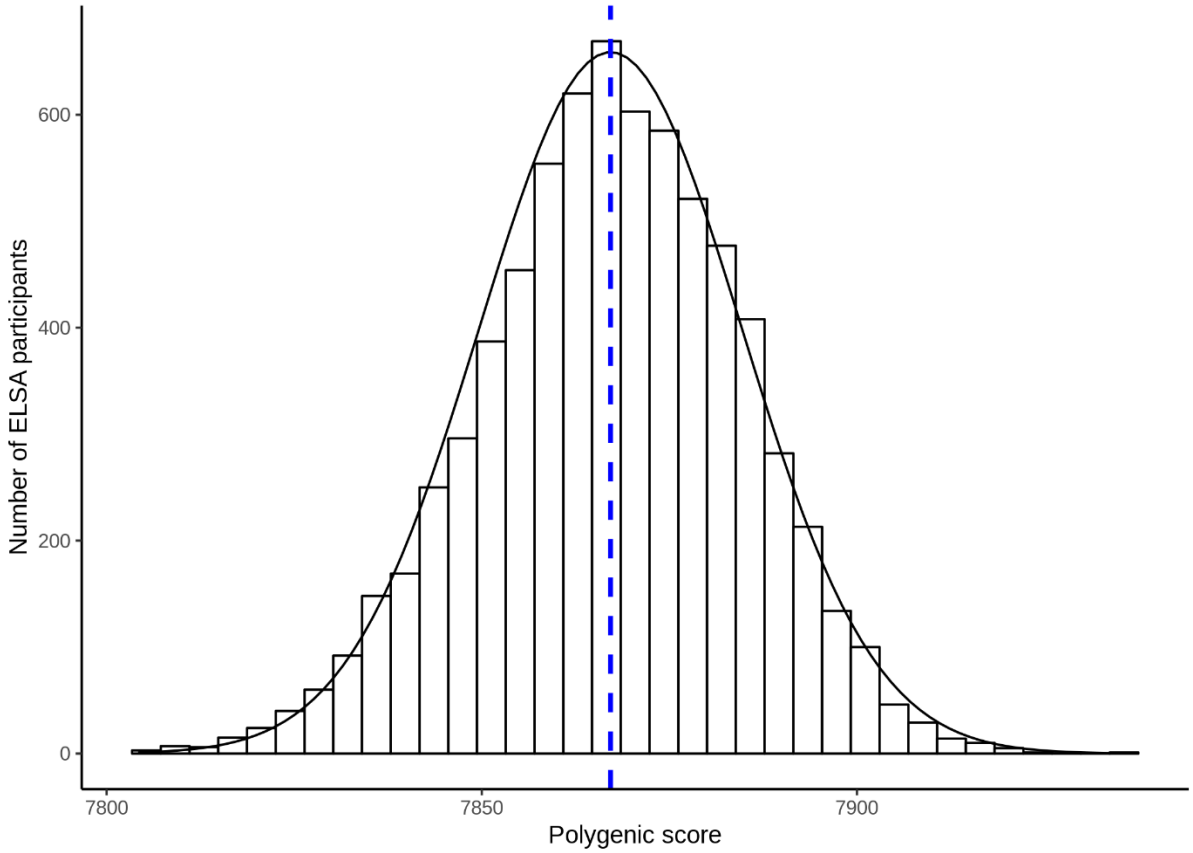
The distribution of PGS for social deprivation in ELSA is depicted in **Figure 5**; the summary statistics for PGS for social deprivation are provided in **Table 3**. The PGS for social deprivation contains 1,341,112 SNPs that overlapped between the ELSA genetic database and the GWAS meta-analysis

Table 3*. The summary statistics for PGS for* Social Deprivation

PGSs	Sample Size	Minimum	Maximum	Range	Median	Mean	SE (mean)
SES	7183	7804.2	7934.5	130.3	7867.5	7867.2	0.21

PGS, polygenic score; SE, standard error; SES,

Figure 5. Distribution of PGS for Social Deprivation



3.3.3. Psychopathology

3.3.3.1. Alzheimer's disease

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 IGAP GWAS meta-analysis files are publicly available on their data download page (Supplementary Table 2).

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.

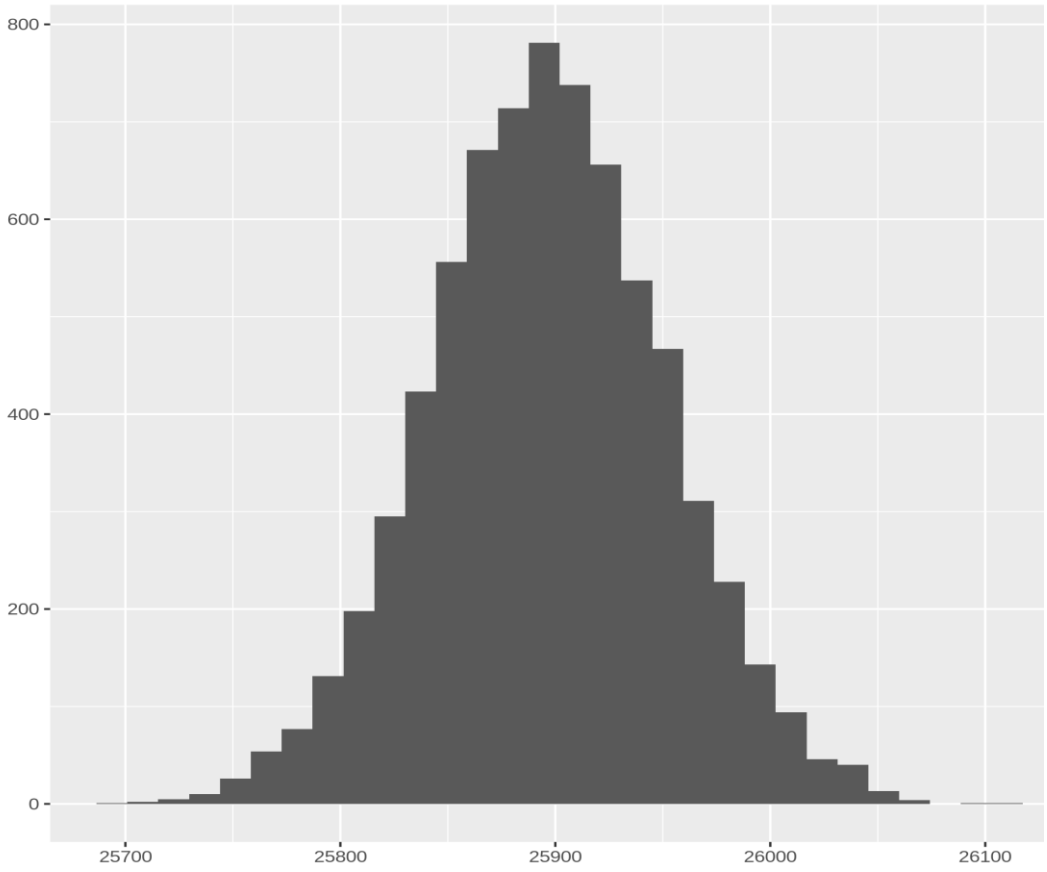
The distribution of PGS for AD in ELSA is depicted in **Figure 6**; the summary statistics for PGS for AD are provided in **Table 4**. The PGS for AD contains 1,191,420 SNPs that overlapped between the ELSA genetic database and the GWAS meta-analysis. It is important to note that the ELSA PGS for AD was created without including the two variants that contribute to ApoE status (rs7412, rs429358).

Table 4*. presents the summary statistics for PGS for Alzheimer's disease*

PGS for AD	Sample Size	Minimum	Maximum	Range	Median	Mean	SE (mean)
	7183	25696	26112.5	416.5	25896	25896.2	0.64

PGS, polygenic score; AD, Alzheimer's disease; SE, standard error

Figure 6. *Distribution of PGS for Alzheimer's disease*



3.3.3.2. 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 GWAS meta-analysis files are publicly available on the SSGAC website which can be found in Supplementary Table 2. The SSGAC GWAS included 180,866 individuals and meta-analysed publicly available results from a study performed by the Psychiatric Genomics Consortium (PGC)[31] (n_{cases}=9,240, $n_{controls}$= 9,519) with results from analyses of UK Biobank (UKB) data[32] (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[32]. 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[31].

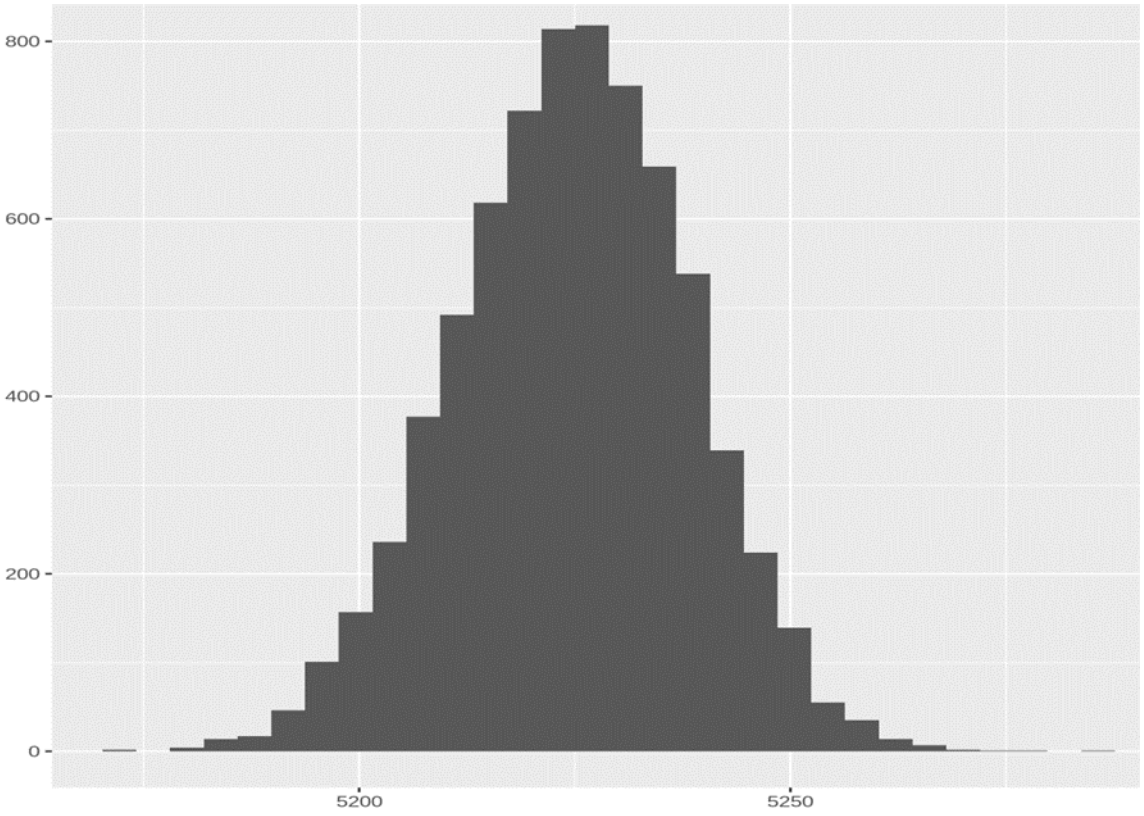
The distribution of PGS for Depressive Symptoms in ELSA is depicted in **Figure 7**; the summary statistics for PGS for Depressive Symptoms are provided in **Table 5**. GWAS summary statistics contained 6,524,474 SNPs; of these, 1,187,563 SNPs overlapped with the ELSA genetic database and were included in the PGS for depressive symptoms phenotype.

Table 5. *The summary statistics for PGS for Depressive Symptoms (DS)*

PGS for DS	Sample Size	Minimum	Maximum	Range	Median	Mean	SE (mean)
	7183	5170.3	5283.8	113.5	5225	5224.7	0.16

PGS, polygenic score; SE, standard error

Figure 7. *Distribution of PGS for Depressive Symptoms*



3.3.3.3. 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[33]. 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.

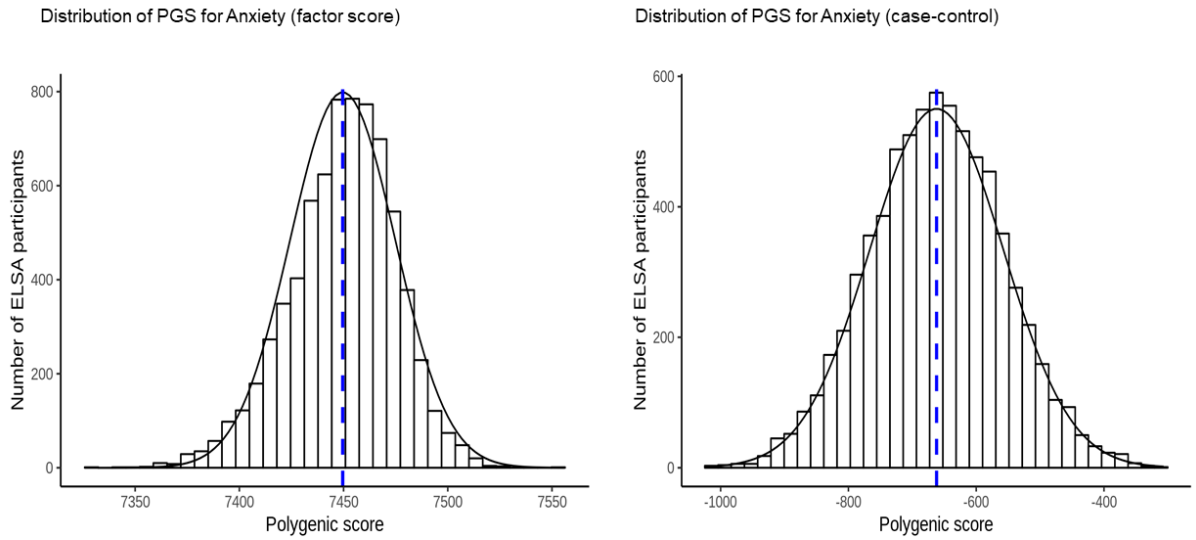
The distribution of PGS for Anxiety (case-control, factor score) in ELSA is depicted in **Figure 8**; the summary statistics for PGS for Anxiety (case-control, factor score) are provided in **Table 6**. From the ANGST meta-analysis, 1,137,311 SNPs overlapped with the ELSA genetic database and were included 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.

Table 6. *The summary statistics for PGS for Anxiety (case-control, factor score)*

PGS	Sample Size	Minimum	Maximum	Range	Median	Mean	SE (mean)
Anxiety (factor score)	7183	7332.1	7556.0	223.9	7451.6	7449.6	0.30
Anxiety (case-control)	7183	-1010.4	-306.6	703.9	-661.2	-662.0	1.23

PGS, polygenic score; SE, standard error

Figure 8. *Distribution of PGS for Anxiety (case-control, factor score)*



3.3.3.4. Insomnia Complaints

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[34]. 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). These GWAS summary statistics are publically available (Supplementary Table 2).

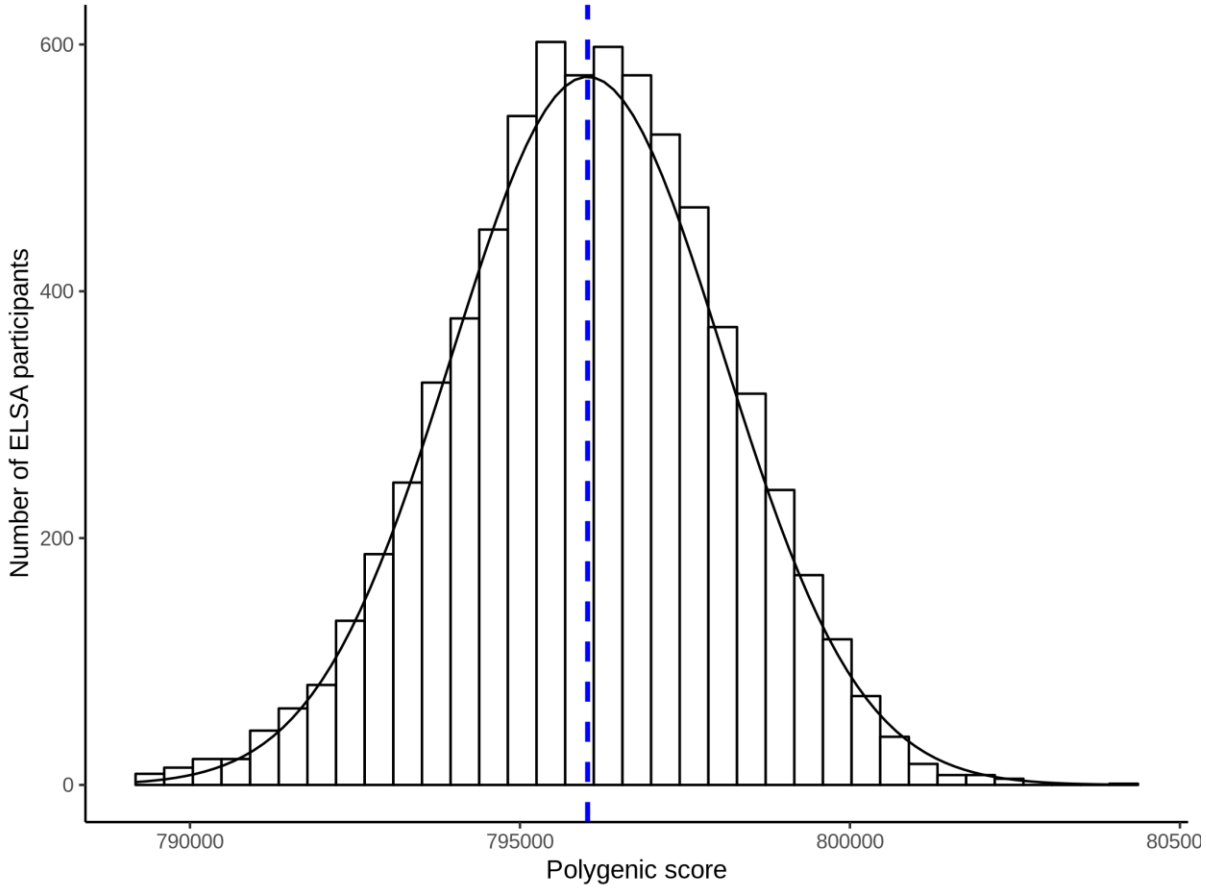
The distribution of PGS for Insomnia Complaints in ELSA is depicted in **Figure 9**; the summary statistics for PGS for Insomnia Complaints are provided in **Table 7**. The PGS contain 803,361 SNPs that overlapped between the ELSA genetic database and the GWAS meta-analysis; these SNPs were included in the PGS for Insomnia Complaints.

Table 7. *The summary statistics for PGS for Insomnia Complaints*

PGS	Sample Size	Minimum	Maximum	Range	Median	Mean	SE (mean)
	7183	789190	803945	14755	796054	796025.2	24.2

PGS, polygenic score; SE, standard error

Figure 9. Distribution of PGS for Insomnia Complaints



3.3.3.5. Schizophrenia (2014)

The PGSs for schizophrenia were created using results from a 2014 GWAS conducted by the Schizophrenia Working Group of the Psychiatric Genomics Consortium (PGC)[35]. The course of these GWAS summary statistics are publically available and provided in Supplementary Table 2. 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.

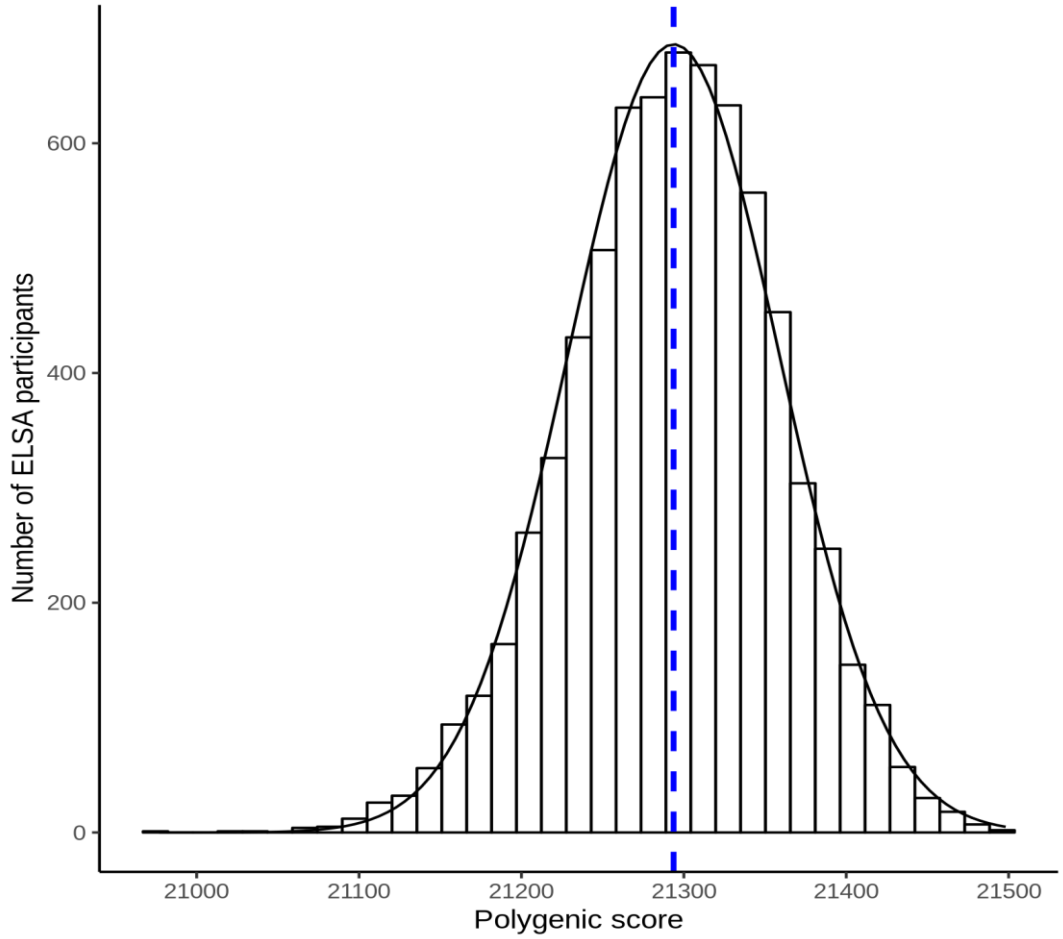
The distribution of PGS for Schizophrenia in ELSA is depicted in **Figure 10**; the summary statistics for PGS for Schizophrenia are provided in **Table 8**. The PGS contain 1,278,742 SNPs that overlapped between the ELSA genetic database and the GWAS meta-analysis; these SNPs were included in the PGS for Schizophrenia.

Table 8*. The summary statistics for PGS for Schizophrenia in ELSA*

PGS	Sample Size	Minimum	Maximum	Range	Median	Mean	SE (mean)
	7183	20976.9	21498.2	521.3	21295.7	21293.7	0.77

PGS, polygenic score; SE, standard error

Figure 10. *Distribution of PGS for Schizophrenia (2014) in ELSA*



3.3.3.6. 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]. These SSGAC GWAS meta-analysis files are publicly available (Supplementary Table 2).

The subjective wellbeing analyses included 298,420 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.

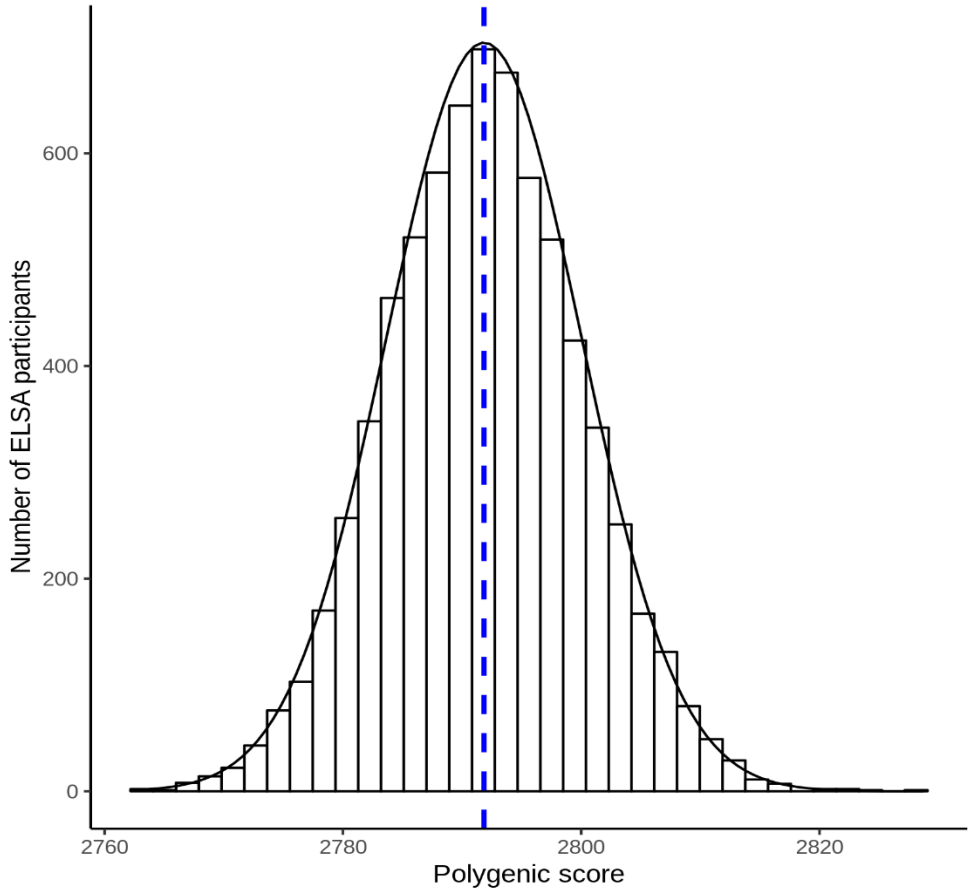
The distribution of PGS for SWB in ELSA is depicted in **Figure 11**; the summary statistics for PGS for SWB are presented in **Table 9**. GWAS summary statistics contained 2,268,674 SNPs; of these, 748,500 SNPs overlapped with the ELSA genetic database and were included in the PGS for SWB phenotype.

Table 9*. The summary statistics for PGS for Subjective Well-Being in ELSA*

PGS	Sample Size	Minimum	Maximum	Range	Median	Mean	SE (mean)
	7183	2763.7	2828.7	6465.0	2791.8	2791.8	0.10

PGS, polygenic score; SE, standard error

Figure 11. *Distribution of PGS for Subjective Well-Being*



3.3.4. Physical health and longevity

3.3.4.1. Coronary Artery Disease

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[36]. The GWAS meta-analysis files are publicly available (Supplementary Table 2).

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 $p < 5 \times 10^{-8}$ 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.

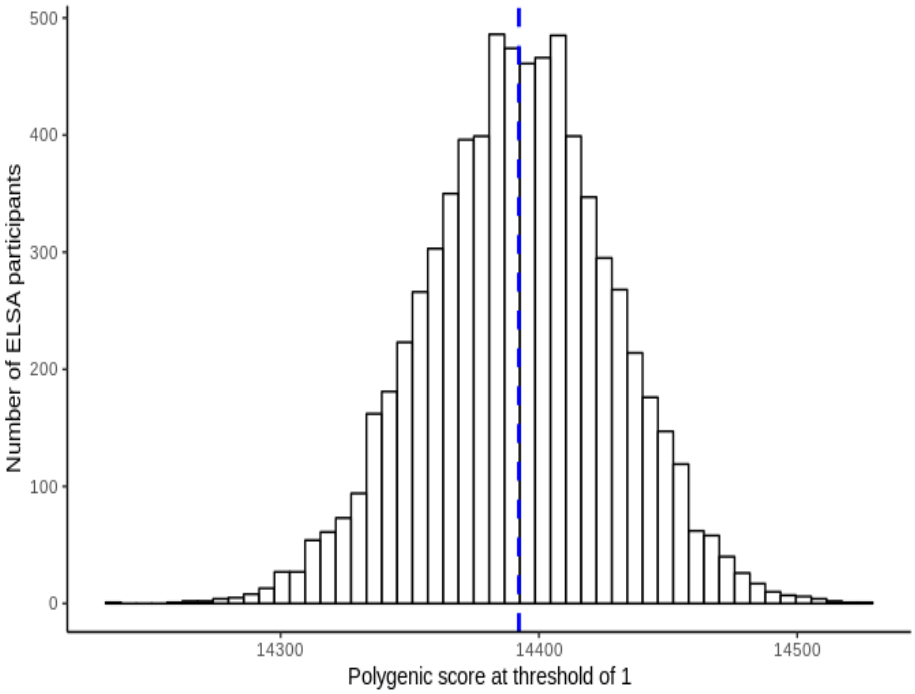
The distribution of PGS for CAD in ELSA is depicted in **Figure 11**; the summary statistics for PGS for CAD are provided in **Table 9**. The PGS contain 783,413 SNPs that overlapped between the ELSA genetic database and the GWAS meta-analysis; these SNPs were included in the PGS for CAD.

Table 10*. The summary statistics for PGS for CAD*

PGS for CAD	Sample Size	Minimum	Maximum	Range	Median	Mean	SE (mean)
	7183	14234.7	14525.7	291	14392.6	14392.3	0.43

PGS, polygenic score; CAD, *Coronary Artery disease*; SE, standard error

Figure 12. *Distribution of PGS for Coronary Artery Disease*



3.3.4.2. Type II Diabetes

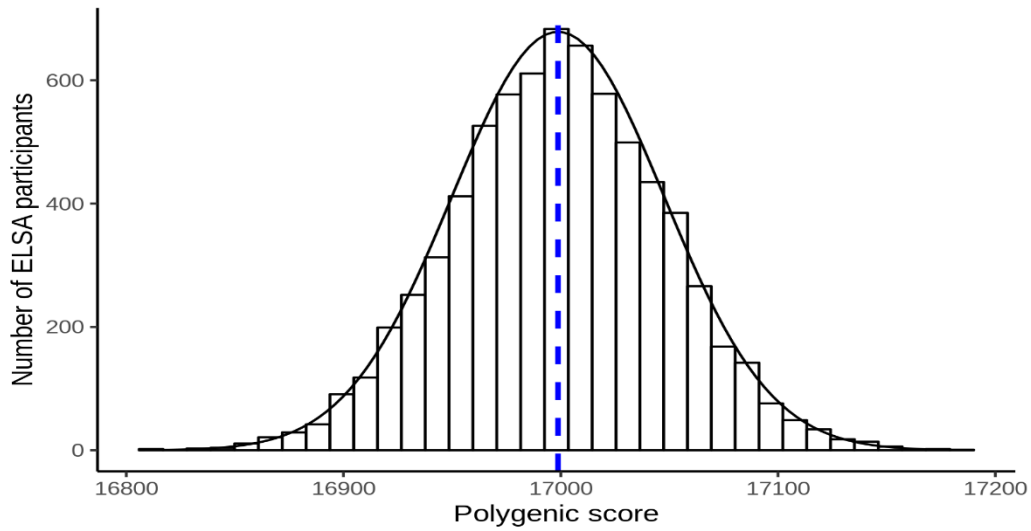
PGSs for Type II Diabetes (T2D) were created using GWAS meta-analysis results from a 2012 study conducted by the DIAbetes Genetics Replication and Meta-analysis (DIAGRAM) Consortium[37]. The GWAS meta-analysis files are publicly available (Supplementary Table 2). The stage one (discovery) meta-analysis 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 22,669 cases and 58,119 controls, including 1,178 cases and 2,472 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 11**; the summary statistics for PGS for T2D are presented in **Table 9**. 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.

***Table 11**. The summary statistics for PGS for T2D*

PGS for T2D	Sample Size	Minimum	Maximum	Range	Median	Mean	SE (mean)
	7183	16806.4	17179.5	373.1	16998.8	16998.7	0.57

PGS, polygenic score; SE, standard error; T2D, Type II diabetes

Figure 13. *Distribution of PGS for Type II Diabetes*



3.3.4.3. General Cognition

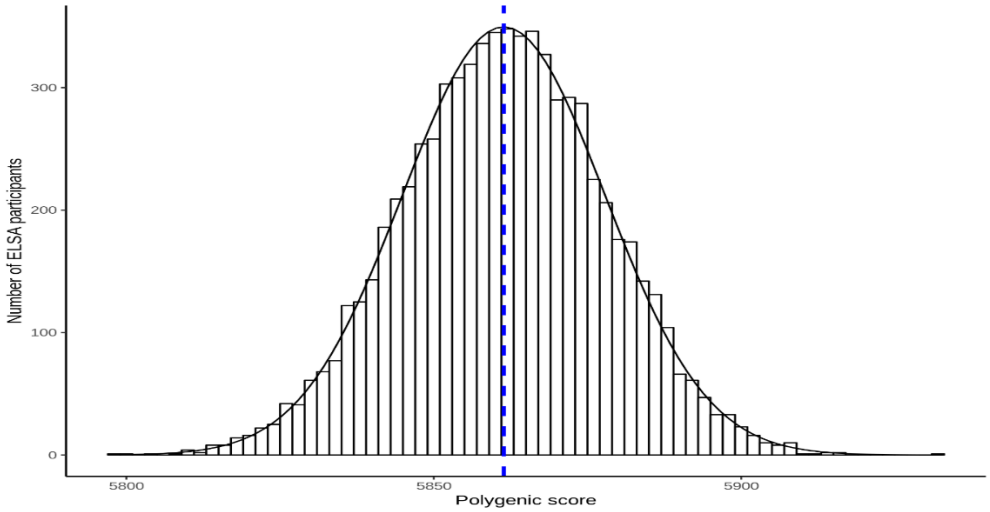
The PGSs for general cognition were created using results from a 2015 GWAS conducted across 31 cohorts by the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) consortium[38]. These CHARGE GWAS meta-analysis files are publicly available (Supplementary Table 2). A total of 53,949 participants 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 12**; the summary statistics for PGS for general cognition are presented in **Table10**. A total of 2,473,946 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.

Table 12*. The summary statistics for PGS for General Cognition*

PGS General Cognition	Sample Size	Minimum	Maximum	Range	Median	Mean	SE (mean)
	7183	5798.4	5932.8	134.3	5861.6	5861.4	0.19

PGS, polygenic score; SE, standard error

Figure 14. *Distribution of PGS for General Cognition*



3.3.4.4. 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 [39]. The summary statistics from this GWAS meta-analysis are freely available online (Supplementary Table 2). 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.

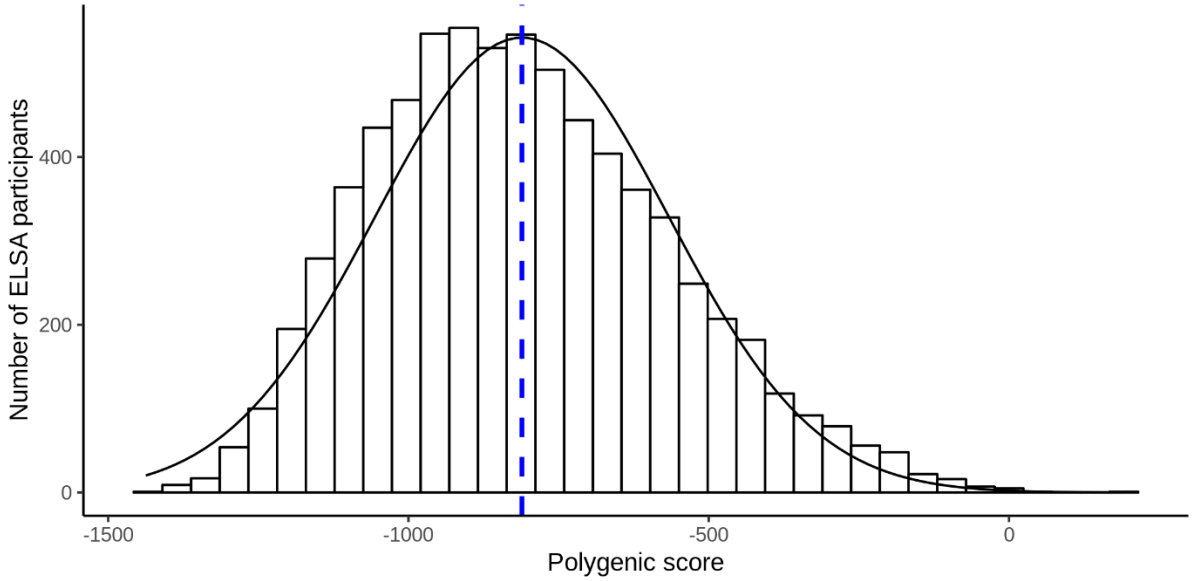
The distribution of PGS for RA in ELSA is depicted in **Figure 13**; the summary statistics for PGS for RA are presented in **Table 11**. A total of 8,747,962 SNPs were included in the meta-analysis summary statistics for RA. Of these, 1,100,616 SNPs overlapped with the ELSA genetic database and were included in the PGS for the Rheumatoid arthritis phenotype.

Table 13*. The summary statistics for PGS for Rheumatoid arthritis*

PGS RA	Sample Size	Minimum	Maximum	Range	Median	Mean	SE (mean)
	7183	-1437.4	187.5	1624.8	-831.7	-811.0	2.87

PGS, polygenic score; SE, standard error; RA, Rheumatoid arthritis

Figure 15. Distribution of PGS for Rheumatoid Arthritis in ELSA



3.3.4.5. 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[40]. The GWAS meta-analysis files are publicly available and can be downloaded from online (Supplementary Table 2).

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 anlayses. 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.

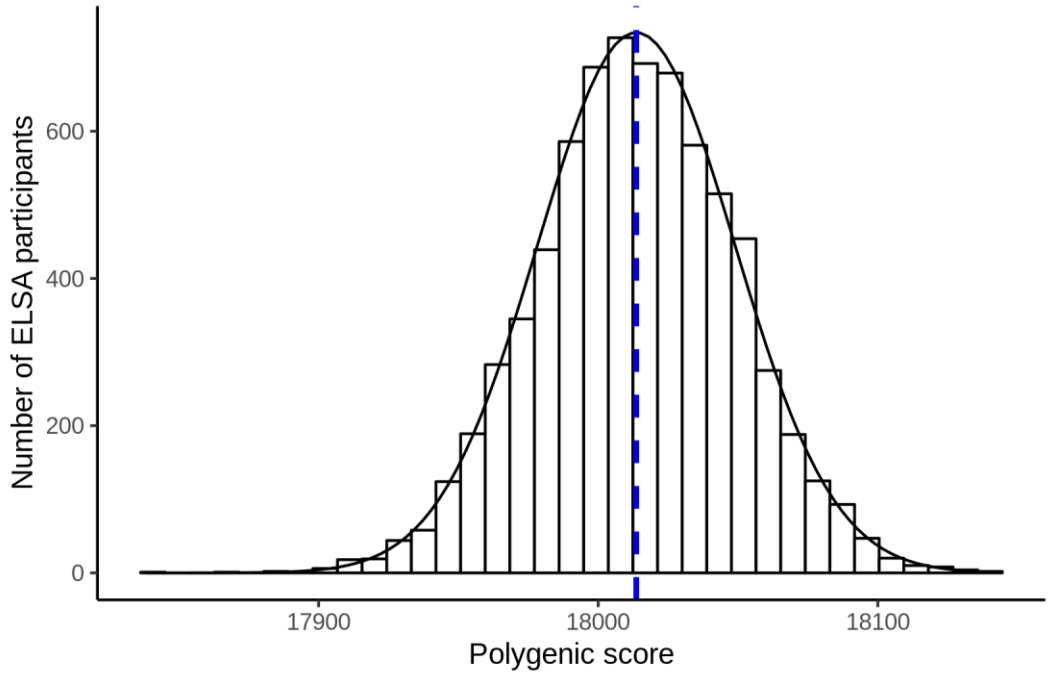
The distribution of PGS for MI in ELSA is depicted in **Figure 14**; the summary statistics for PGS for MI are presented in **Table 12**. The European ancestry PGSs contain 1,299,282 SNPs that overlapped between the ELSA genetic database and the GWAS meta-analysis; these SNPs were included in the PGS.

Table 14*. The summary statistics for PGS for Myocardial infarction*

PGS MI	Sample Size	Minimum	Maximum	Range	Median	Mean	SE (mean)
	7183	17838.4	18137.8	299.4	18013.4	18013.6	0.42

PGS, polygenic score; SE, standard error; MI, Myocardial infarction

Figure 16. Distribution of PGS for Myocardial Infarction in ELSA



3.3.4.6. Longevity

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[41]. The GWAS summary statistics for this phenotype were obtained from the GRASP (Genome-Wide Repository of Associations Between SNPs and Phenotypes) [42] which is publicly available, the link to which can be found in Supplementary Table 2.

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.

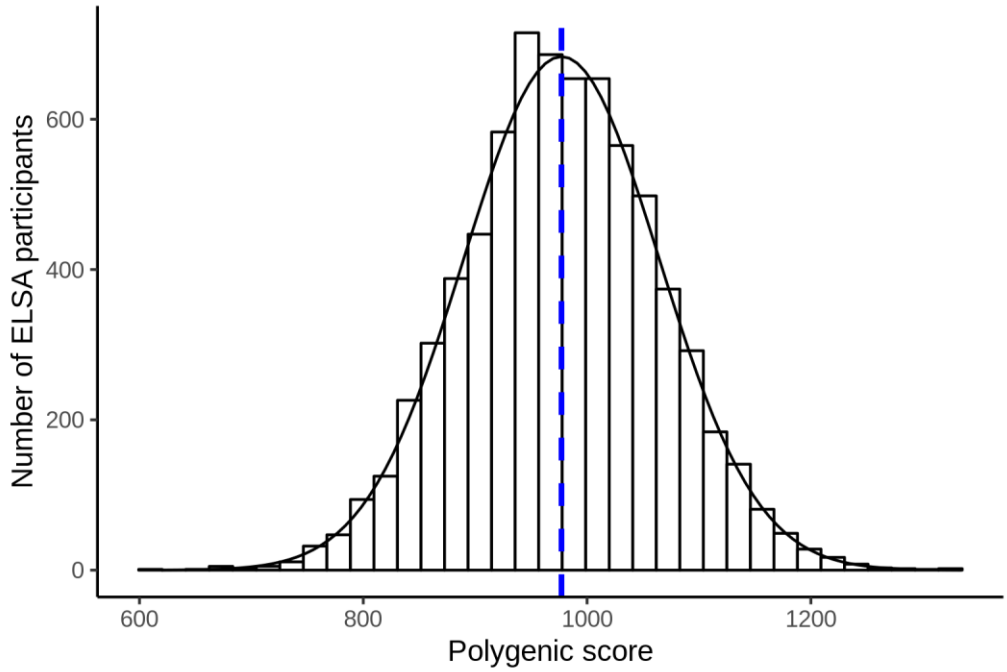
The distribution of PGS for longevity in ELSA is depicted in **Figure 15**; the summary statistics for PGS for longevity are presented in **Table 13**. A total of 2,588,525 SNPs were included in the summary statistics. Of these, 757,472 SNPs overlapped with the ELSA genetic database and were included in the PGS for this phenotype.

Table 15*. The summary statistics for PGS for Longevity*

PGS Longevity	Sample Size	Minimum	Maximum	Range	Median	Mean	SE (mean)
	7183	619.4	1334.4	714.9	976.5	977.1	1.03

PGS, polygenic score; SE, standard error

Figure 17. *Distribution of PGS for Longevity in ELSA*



3.3.4.7. Sleep Duration

PGSs for sleep duration in ELSA using the GWAS summary statistics perfumed using the data from the UK Biobank[43]. 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. Summary GWAS statistics is made available at the UK Biobank website (Supplementary Table 2).

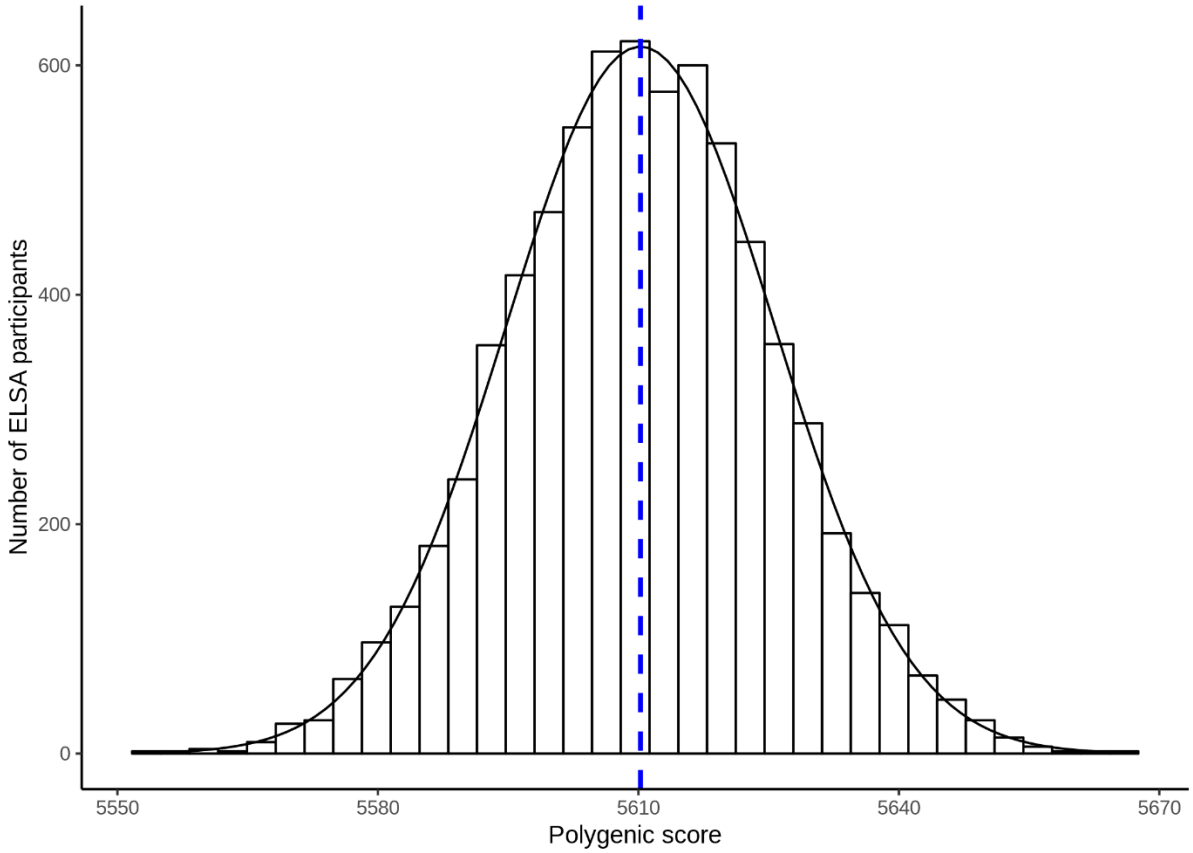
The distribution of PGS for Sleep Duration in ELSA is depicted in **Figure 16**; the summary statistics for PGS for Sleep Duration are presented in **Table 14**. A total of 948,331 SNPs overlapped with the ELSA genetic database with the GWAS summary statistics and were included in the PGS for this phenotype.

Table 16*. The summary statistics for PGS for Sleep Duration*

PGS Sleep Duration	Sample Size	Minimum	Maximum	Range	Median	Mean	SE (mean)
	7183	5552.6	5665.2	112.6	5610.3	5610.2	0.18

PGS, polygenic score; SE, standard error

Figure 18. Distribution of PGS for Sleep Duration



3.3.5. Anthropomorphic traits

3.3.5.1. Height

PGS for height was created using the results from a 2014 study conducted by the Genetic Investigation of ANthropometric Traits (GIANT) consortium [44]. The GWAS meta-analysis files are publicly available on their data download page (Supplementary Table 2). 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.

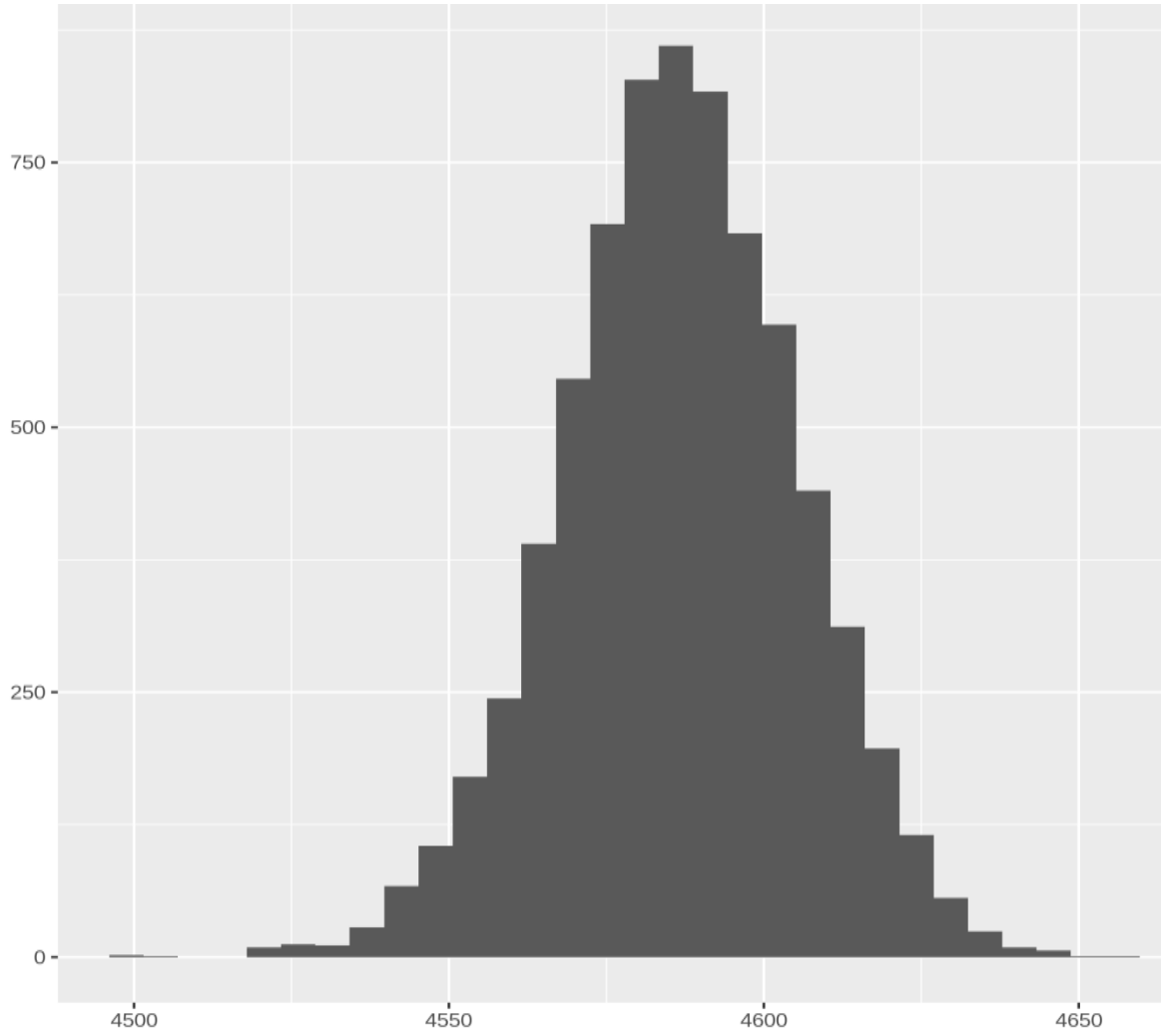
The distribution of PGS for Height in ELSA is depicted in **Figure 17**; the summary statistics for PGS for Height are provided in **Table 15**. The PGS contains 831,045 SNPs that overlapped between the ELSA genetic database and the GIANT GWAS meta-analysis and that were included PGS for this phenotype.

Table 17*. The summary statistics for PGS for Height*

PGS for Height	Sample Size	Minimum	Maximum	Range	Median	Mean	SE (mean)
	7183	4498.4	4656.5	158.1	4586.6	4586.5	0.22

PGS, polygenic score; SE, standard error

Figure 19. *Distribution of PGS for Height*



3.3.5.2. Body Mass Index (BMI)

PGS for BMI was created using results from a 2015 GWAS conducted by the Genetic Investigation of ANthropometric Traits (GIANT) consortium [45]. The GIANT GWAS meta-analysis files are publicly available (Supplementary Table 2). 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.

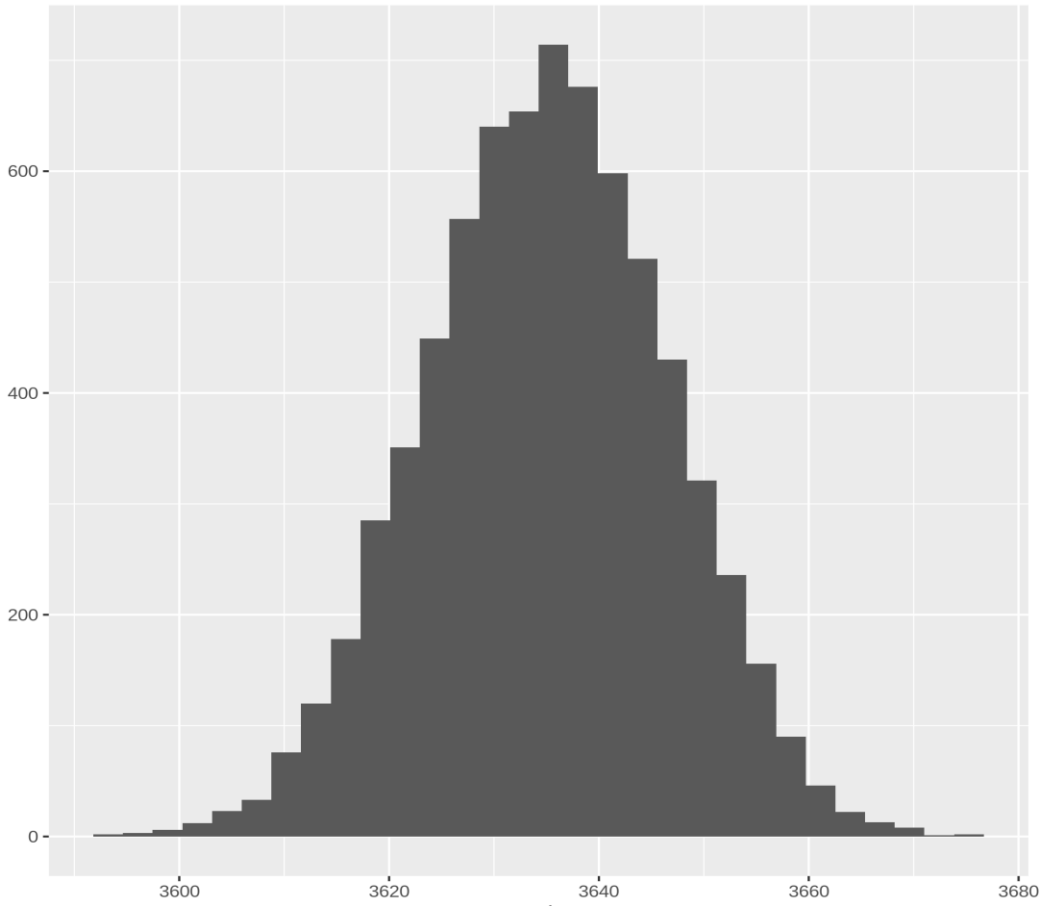
The distribution of PGS for BMI in ELSA is depicted in **Figure 15**; the summary statistics for PGS for BMI are provided in **Table 18**. The PGS contains 795,650 SNPs that overlapped between the ELSA genetic database and the GIANT GWAS meta-analysis which were included in PGS.

Table 18*. The summary statistics for PGS for BMI*

PGS for BMI	Sample Size	Minimum	Maximum	Range	Median	Mean	SE (mean)
	7183	3594.0	3676	82.1	3635.1	3635	0.14

PGS, polygenic score; BMI, Body Mass Index; SE, standard error

Figure 20. *Distribution of PGS for BMI*



3.3.5.3. Waist circumference & Waist-Hip Ratio

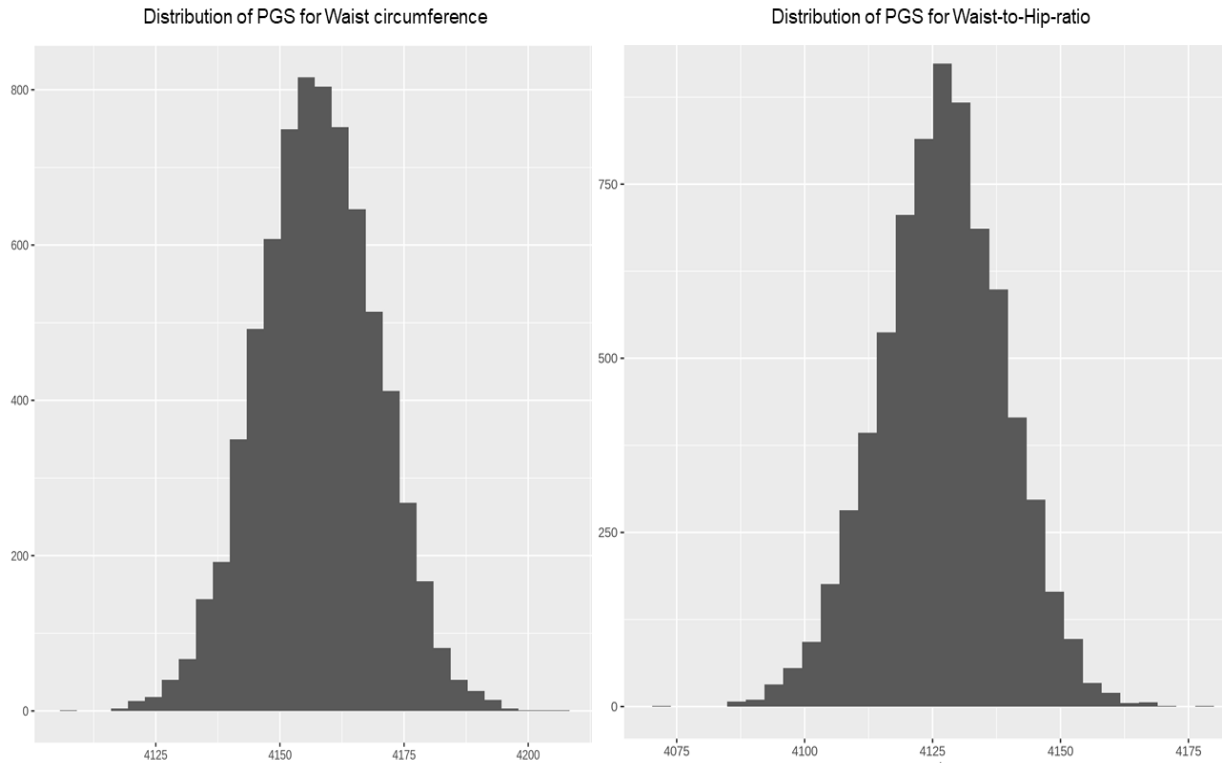
PGS for waist circumference (WC) and waist-to-hip ratio (WHR) were created using results from a 2015 study conducted by the Genetic Investigation of ANthropometric Traits (GIANT) consortium [46]. The GWAS meta-analysis files are publicly available from the webpage (Supplementary Table 2). GWAS meta-analysis was performed on a sample of 142,762 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 93,057 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 19**; the summary statistics for PGS for WC and WHR are provided in **Table 17**. 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.

Table 19*. The summary statistics for PGSs for WC and WHR*

	Sample Size	Minimum	Maximum	Range	Median	Mean	SE (mean)
PGS for WC	7183	4106.3	4205.5	99.2	4157.5	4157.6	0.14
PGS for WHR	7183	4070.5	4176.6	106.1	4127.1	4126.9	0.14

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

Figure 21. *Distribution of PGS for WC and WHR*



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

3.3.6. Behavioural traits

3.3.6.1. Smoking behaviour

PGS for smoking behaviours in ELSA was contracted using the results from the Tobacco and Genetics (TAG) Consortium (2010) [47]. 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 GWAS meta-analysis files for this phenotype are publicly available; the link to the website and the file can be found in Supplementary Table 2.

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.

3.3.6.1.1. Number of cigarettes smoked per day

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 case-control in 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 20**; the summary statistics for PGS for CPD are provided in **Table 18**. 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.

3.3.6.1.2. Smoking initiation (ever/never)

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 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.

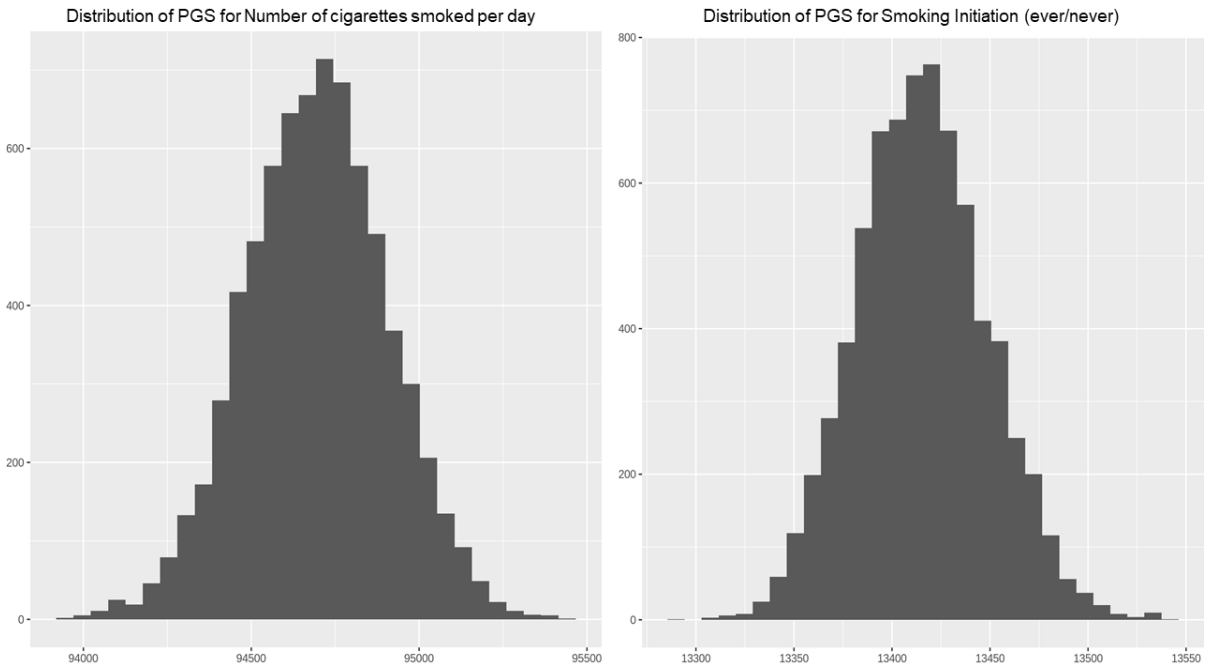
The distribution of PGS for Smoking initiation in ELSA is depicted in **Figure 20**; the summary statistics for PGS for smoking initiation are provided in **Table 18**. The TAG GWAS summary statistics for this smoking phenotype was based on the sample of 143,023 individuals and contained 2,455,846 SNPs; of these, 804,337 SNPs overlapped with the ELSA genetic database and were included in the PGS for smoking initiation phenotype.

***Table 20**. The summary statistics for PGS for two smoking behaviour*s

PGS	Sample Size	Minimum	Maximum	Range	Median	Mean	SE (mean)
CPD	7183	93951.2	95447.3	1496.1	94696.5	94694.2	2.49
Smoking initiation	7183	13289.0	13540.9	251.9	13414.7	13415.3	0.39

PGS, polygenic score; CPD, number of cigarettes smoked per day; SE, standard error

Figure 22. *Distribution of PGSs for smoking behaviours*



3.3.6.1.3. Daily Alcohol Intake

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[48]. These GWAS summary statistics are publicly available; the link to the website and the file can be found in Supplementary Table 2. 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.

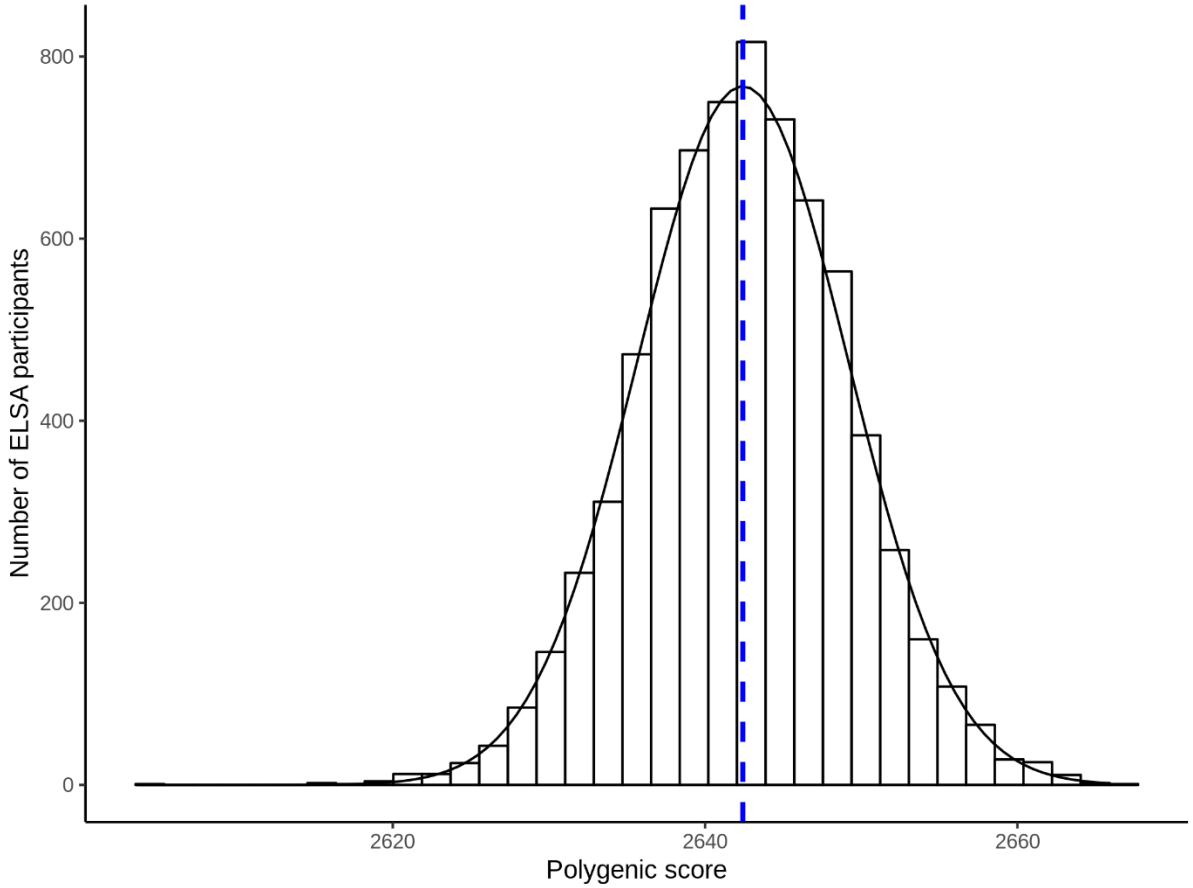
The distribution of PGS for Daily Alcohol Intake in ELSA is depicted in **Figure 21**; the summary statistics for PGS for Daily Alcohol Intake are provided in **Table 19**. The GWAS summary statistics for the Daily Alcohol Intake phenotype included 2,462,742 SNPs; of these, 800,524 SNPs overlapped with the ELSA genetic database and were included in the PGS for Daily Alcohol Intake phenotype.

Table 21*. Summary statistics for PGS for Daily Alcohol Intake (in grams of alcohol per day)*

PGS	Sample Size	Minimum	Maximum	Range	Median	Mean	SE (mean)
DAI	7183	2603.6	2666.0	62.4	2642.5	2642.4	0.08

PGS, polygenic score; DAI, daily alcohol intake; SE, standard error

Figure 23. *Distribution of PGS for Daily Alcohol Intake (in grams of alcohol per day)*



3.3.7. Biological outcomes

3.3.7.1. Morning Plasma cortisol

PGS for smoking behaviours 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 [49]. 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 summary statistics from the CORNET GWAS meta-analysis are publicly available; the link to the website and the file can be found in Supplementary Table 2.

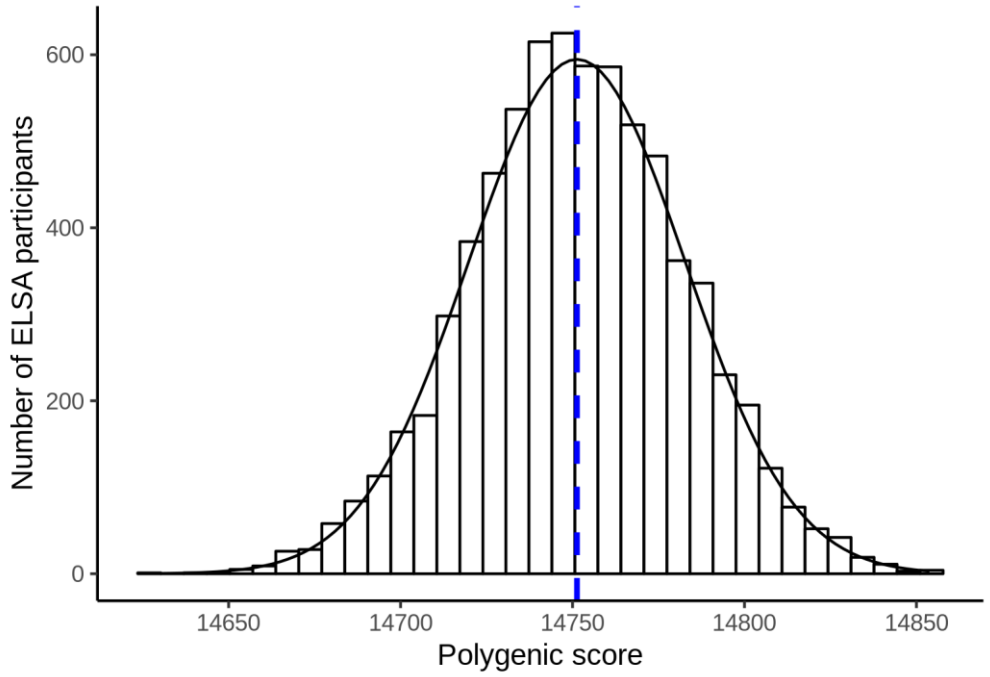
The distribution of PGS for Morning Plasma Cortisol in ELSA is depicted in **Figure 22**; the summary statistics for PGS for Morning Plasma Cortisol are provided in **Table 20**. The CORNET GWAS summary statistics for this phenotype contained 2,660,191 SNPs; of these, 837,709 SNPs overlapped with the ELSA genetic database and were included in the PGS for Morning Plasma Cortisol phenotype.

Table 22*. The summary statistics for PGS for Morning Plasma cortisol*

PGS	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

PGS, polygenic score; CPD, number of cigarettes smoked per day; SE, standard error

Figure 24. *Distribution of PGSs for Morning Plasma cortisol*



3.3.8. Reproductive behaviour

3.3.8.1 Age at Menarche

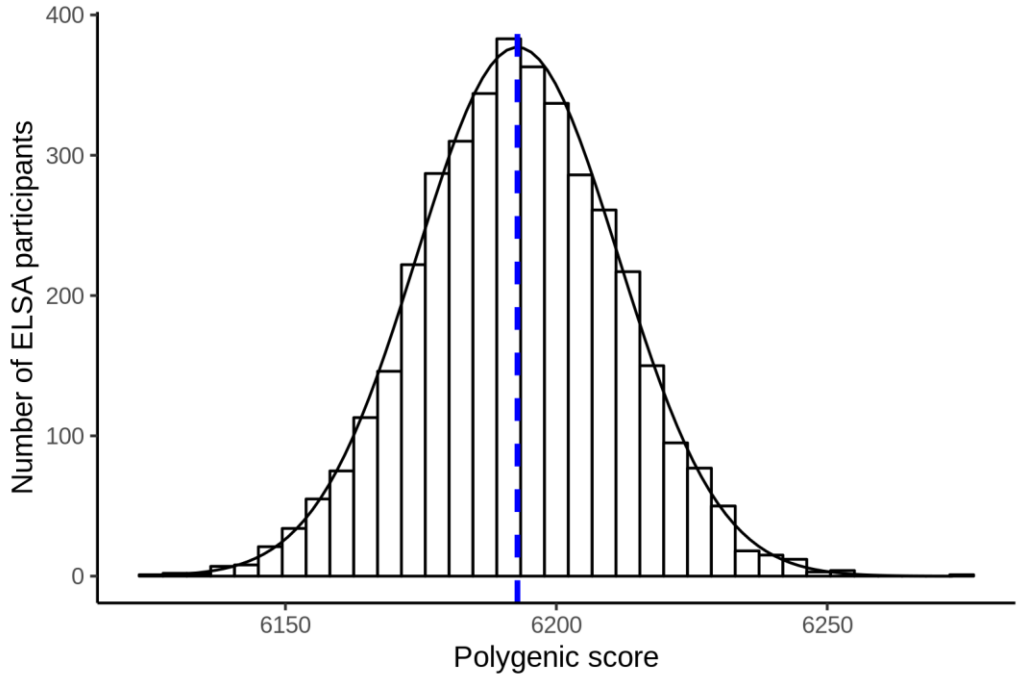
PGSs for age at menarche were created using results from a 2014 study conducted by the Reproductive Genetics (ReproGen) consortium[50]. The GWAS meta-analysis files are publicly available on the ReproGen data download page (Supplementary Table 2). The ReproGen meta-analysis included 182,416 women of 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 distribution of PGS for Age at Menarche in ELSA is depicted in **Figure 23**; the summary statistics for PGS for Age at Menarche are provided in **Table 21**. 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.

***Table 23**. The summary statistics for PGS for Age at Menarche*

PGS	Sample Size	Minimum	Maximum	Range	Median	Mean	SE (mean)
Age at Menarche	3878	6123.6	6273.2	149.6	6192.6	6192.8	0.30

PGS, polygenic score; CPD, number of cigarettes smoked per day; SE, standard error

Figure 25. *Distribution of PGSs for Age at Menarche*



3.3.8.2. Age at Menopause

PGSs for age at menarche were created using results from a 2014 study conducted by the Reproductive Genetics (ReproGen) consortium[50]. The GWAS meta-analysis files are publicly available on the ReproGen data download page (Supplementary Table 2).

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.

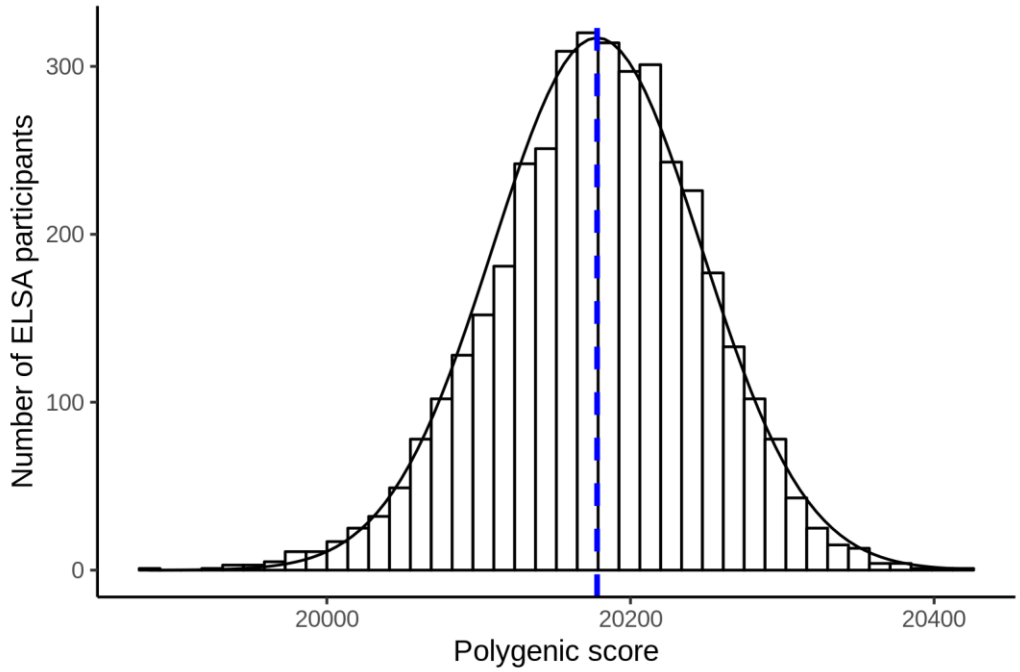
The distribution of PGS for Age at Menopause in ELSA is depicted in **Figure 24**; the summary statistics for PGS for Age at Menopause are provided in **Table 22**. The PGSs contain 777,339 SNPs that overlapped between the ELSA genetic database and the GWAS meta-analysis; these SNPs were included in the PGS for this phenotype.

Table 24*. The summary statistics for PGS for Age at Menopause*

PGS	Sample Size	Minimum	Maximum	Range	Median	Mean	SE (mean)
Age at Menopause	3878	19882.8	20418.7	535.9	20179.8	20178.0	1.10

PGS, polygenic score; CPD, number of cigarettes smoked per day; SE, standard error

Figure 26. *Distribution of PGSs for Age at Menopause*



3.3.8.3. Age at first birth – Female & Male

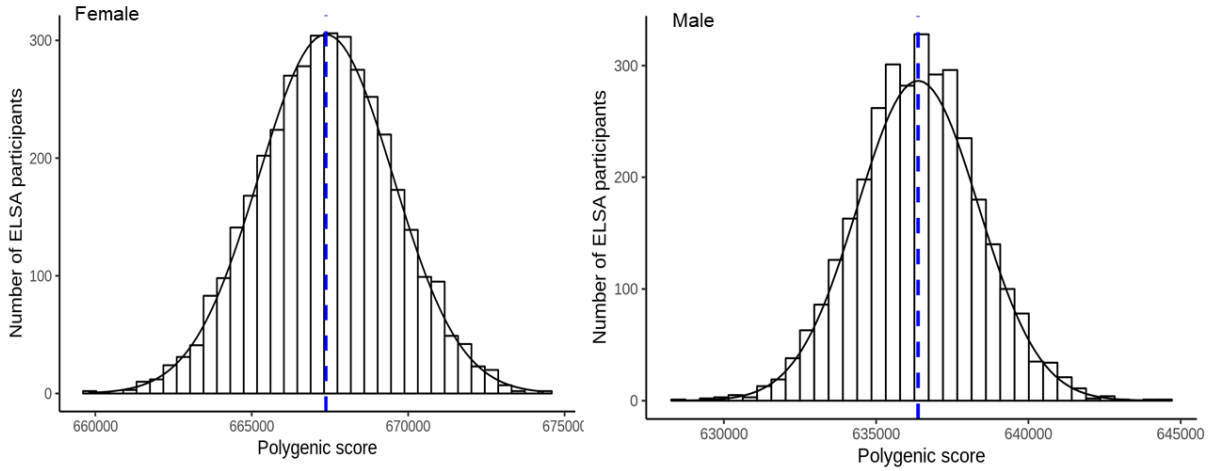
PGSs for the Age at First Birth (AFB) for women and men were created using the GWAS summary statistics conducted by Barban et al. (2016)[51]. The GWAS meta-analysis files are publicly available (Supplementary Table 2). The total sample size of the meta-analysis for AFB was *n*=251,151. Cohorts uploaded 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 distribution of PGS for AFB in ELSA is depicted in **Figure 25**; the summary statistics for PGS for AFB are provided in **Table 23**. 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.

Table 25*. The summary statistics for PGS for Age at first birth: Female and Male*

PGS Age at first birth	Sample Size	Minimum	Maximum	Range	Median	Mean	SE (mean)
Female	3878	659778	674336	14558	667389	667371.7	34.31
Male	3305	628409	644380	15971	636385	636376.4	34.53

PGS, polygenic score; CPD, number of cigarettes smoked per day; SE, standard error

Figure 27. *Distribution of PGS for Age at first birth – Female & Male*



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

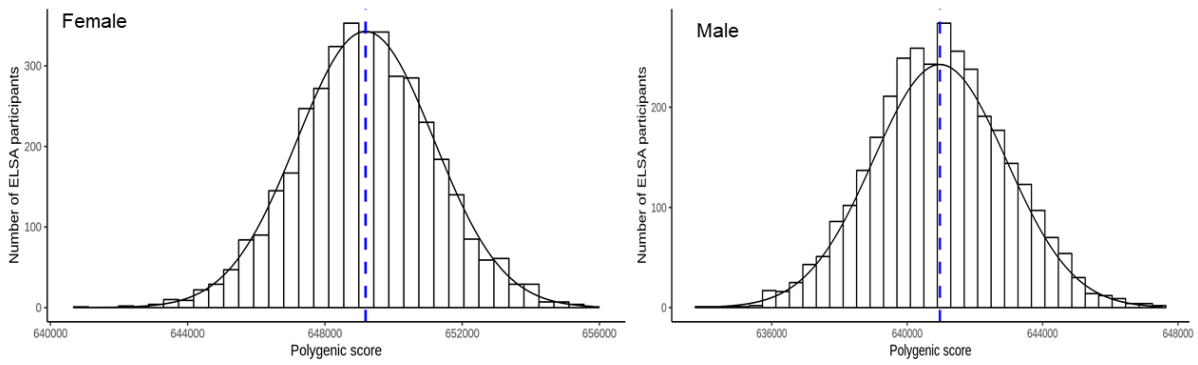
PGSs for the number of children ever born (NEB) for women and men were created using the GWAS summary statistics conducted by Barban et al. (2016)[51]. The GWAS meta-analysis files are publicly available (Supplementary Table 2). The total sample size of the meta-analysis was N=343,072 for NEB pooled. 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 distribution of PGS for NEB in ELSA is depicted in **Figure 26**; the summary statistics for PGS for NEB are provided in **Table 24**. The PGS for NEB for female participants contain 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.

Table 26*. The summary statistics for PGS for Number of children ever born: Female and Male*

PGS number of children ever born	Sample Size	Minimum	Maximum	Range	Median	Mean	SE (mean)
Female	3878	640948	655822	14874	649140	649185.0	31.96
Male	3305	634040	647545	13505	640960	640967.4	34.10

PGS, polygenic score; CPD, number of cigarettes smoked per day; SE, standard error

Figure 28*. Distribution of PGSs for Number of children ever born – Female & Male*



4. SET UP

4.1. Download the PGSs in ELSA

By downloading this freely provided data set, you agree to use its 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_APR_2019.dta
2. List_PGS_SCORES_ELSA_APR_2019.sav
3. List_PGS_SCORES_ELSA_APR_2019.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 to check 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 ELSAAPR 2019.dta
2. Principal Components ELSAAPR 2019.sav
3. Principal Components ELSAAPR 2019.xlsx

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

4.3. Data dictionary

In the data files, the names of the phenotypes for which the PGSs are available are abbreviated. The explanations of the abbreviations are provided in the **Table 27**.

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. 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

Table 27. *Explanations of the abbreviations used in the ELSA_PGS_SCORE files.*

Abbreviations	Explanations of the abbreviations
Age_Birth_F	Age at first birth – Female
Age_Birth_M	Age at first birth –Male
AGE_MENARCHE	Age at Menarche
AGE_MENOPAUS	Age at Menopause
AGREE	Agreeableness
ALZ	Alzheimer's disease
ANXIETY_CC	Anxiety Disorders (case-control)
ANXIETY_FC	Anxiety Disorders (factor score)
BMI	Body Mass Index
CAD	Coronary Artery Disease
CON	Conscientiousness
DAI	Daily Alcohol Intake
DIAB	Type II Diabetes
DS	Depressive Symptoms
EA_2	Educational Attainment - 2
EA_3	Educational Attainment - 3
EXTRA	Extraversion
GC	General Cognitive Functioning
Height	Height
INS_COM	Insomnia Complaints
LONGEVITY	Longevity
M_Plasma	Morning Plasma cortisol
MI	Myocardial infarction
NEB_F	Number of children ever born (NEB) – Female
NEB_M	Number of children ever born (NEB) –Male
NEURO	Neuroticism
OPEN	Openness to Experience
RA	Rheumatoid Arthritis
SEC_DEP	Social Deprivation
SLP_DR	Sleep Duration
SMK_EVER	Smoking initiation (ever/never)
SMK_NUMBER	Number of cigarettes smoked per day
SWB	Subjective Well-Being
SZC	Schizophrenia
Waist	Waist
WHR	Waist-Hip Ratio

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. Ripke, S., et al., *A mega-analysis of genome-wide association studies for major depressive disorder.* Mol Psychiatry, 2013. **18**(4): p. 497-511.
32. 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.
33. Otowa, T., et al., *Meta-analysis of genome-wide association studies of anxiety disorders.* Mol Psychiatry, 2016. **21**(10): p. 1485.
34. 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.
35. *Biological insights from 108 schizophrenia-associated genetic loci.* Nature, 2014. **511**(7510): p. 421-7.
36. 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.
37. 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.
38. 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.
39. Okada, Y., et al., *Genetics of rheumatoid arthritis contributes to biology and drug discovery.* Nature, 2014. **506**(7488): p. 376-81.
40. Consortium, C.D., *A comprehensive 1000 Genomes–based genome-wide association meta-analysis of coronary artery disease.* Nature, 2015. **47**: p. 1121–1130.
41. 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.
42. 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.
43. 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.
44. 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.
45. Locke, A.E., et al., *Genetic studies of body mass index yield new insights for obesity biology.* Nature, 2015. **518**(7538): p. 197-206.
46. Shungin, D., et al., *New genetic loci link adipose and insulin biology to body fat distribution.* Nature, 2015. **518**(7538): p. 187-196.

47. *Genome-wide meta-analyses identify multiple loci associated with smoking behavior.* Nat Genet, 2010. **42**(5): p. 441-7.
48. 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.
49. 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.
50. 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.
51. Barban, N., et al., *Genome-wide analysis identifies 12 loci influencing human reproductive behavior.* Nat Genet, 2016. **48**(12): p. 1462-1472.

6. SUPPLEMENTARY MATERIAL	PAGE
Supplementary Table 1. provides 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.	74
Supplementary Table 2. Outlines details of the GWAS summary statistics used for the phenotypes described in this document	75
Supplementary Figure 1. depicts distribution of 10 principal components once 65 individuals with ancestral admixture were removed from the sample.	78

Supplementary Table 1. *provides 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	
Total removed	*229*	*3.09*
Total remaining	*7183*	*96.91*

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

Supplementary Table 2. Outlines details of the GWAS summary statistics used for the phenotypes described in this document

Phenotype	Consortium	GWAS SNPs	Overlapping with ELSA	GWAS meta-analysis citation	Source of base data
Psychosocial					
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]	https://www.thessgac.org/data (https://www.dropbox.com/s/ho58e9jmytmpaf8/GWAS_EA_excl23andMe.txt?dl=0)
Social Deprivation	-	15,732,391	1,341,112	Hill et al (2016) [29]	https://grasp.nhlbi.nih.gov/FullResults.aspx
Personally types					
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	de Moor et al. (2012)[25]	http://www.tweelingenregister.org/GPC
Openness	GPC	2,305,738	750,564		
Conscientiousness	GPC	2,305,682	750,990		
Psychopathology					
Alzheimer's disease	IGAP	7,055,881	1,191,420	Lambert et al. (2013)[30]	http://web.pasteur-lille.fr/en/recherche/u744/igap/igap_download.php
Depressive symptoms	SSCAG	6,524,474	1,187,563	Okbay e al. (2016)[26]	https://www.thessgac.org/data
Anxiety Disorders (factor score)	ANGST	6,306,612	1,137,311	Otowa et al. (2016)[33]	https://www.med.unc.edu/pgc/results-and-downloads
Anxiety Disorders (case-control)	ANGST		1,068,194	Otowa et al. (2016)[33]	https://www.med.unc.edu/pgc/results-and-downloads
Insomnia Complaints	-	12,444,915	803,361	Hammerschlag et al (2017)[34]	http://ctg.cncr.nl/software/summary_statistics
Schizophrenia (2014)	PGC	9,444,230	1,278,742	Ripke et al. (2014)[35]	https://www.med.unc.edu/pgc/results-and-downloads (scz2.snp.results.txt.gz)
Subjective Well-Being	SSCAG	2,268,674	748,500	Okbay e al (2016)[26]	Provided by request
Behavioural traits					
Smoking Initiation (ever/never)	TAG	2,455,846	804,337	Tobacco and Genetics Consortium (2010) [47]	https://www.med.unc.edu/pgc/results-and-downloads (tag.evrsmk.tbl.gz)
Number of cigarettes smoked per day	TAG	2,459,118	803,092		https://www.med.unc.edu/pgc/results-and-downloads (tag.cpd.tbl.gz)
Daily Alcohol Intake	-	2,462,742	800,524	Schumann et al (2016)[48]	https://grasp.nhlbi.nih.gov/FullResults.aspx

Physical health & Longevity					
Coronary Artery Disease	CARDIoGRAM	2,420,360	783,413	Schunkert et al. (2011)[36]	www.cardiogramplusc4d.org (cad.add.160614.website.txt)
Type II Diabetes	DIAGRAM	2,473,441	761,488	Morris et al. (2012)[37]	http://www.diagram-consortium.org/downloads.html (DIAGRAMv3.2012DEC17.txt).
General cognitive function	CHARGE	2,473,946	795,327	Davies et al. (2015)[38]	https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000930.v6.p1
Rheumatoid arthritis		8,747,962	1,100,616	Okada et al. (2014)[39]	http://plaza.umin.ac.jp/~yokada/datasource/software.htm
Myocardial infarction	CARDIoGRAM	9,289,491	1,299,282	CARDIoGRAMplusC4D Consortium. (2015)[40]	www.cardiogramplusc4d.org (mi.add.030315.website.txt)
Longevity	CHARGE	2,588,525	757,472	Broer et al. (2014)[41]	https://grasp.nhlbi.nih.gov/FullResults.aspx
Sleep Duration	-	32,449,020	948,331	Lane et al (2017)[43]	http://biobank.ctsu.ox.ac.uk/
Anthropomorphic traits					
Body Mass Index	GIANT	2,554,623	795,650	Locke et al. (2015)[45]	https://www.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium_data_files
Height	GIANT	2,550,858	831,045	Wood et al. (2014)[44]	https://www.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium_data_files
Waist circumference	GIANT	2,565,407	801,114	Shungin et al. (2015)[46]	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
Biological outcomes					
Plasma cortisol (morning)	CORNET	2,660,191	837,709	Bolton et al. (2014)[49]	https://datashare.is.ed.ac.uk/handle/10283/2787
Reproductive behaviour					
Age at Menarche	ReproGen	2,441,815	793,272	Perry et al. (2014)[50]	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)[51]	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/

IGAP, International Genomics of Alzheimer's Project; GIANT, Genetic Investigation of ANthropometric Traits; CARDIoGRAM, Coronary ARtery DIsease Genome wide Replication and Meta-analysis; SSCAG, Social Science Genetic Association Consortium; TAG, Tobacco and Genetics; ELSA, English Longitudinal Study Of Ageing; DIAGRAM, DIAbetes Genetics Replication and Meta-analysis Consortium; CHARGE, Heart and

Aging Research in Genomic Epidemiology consortium; GPC, Genetics of Personality Consortium; ReproGe; Reproductive Genetics consortium; CORNET, CORtisol NETwork consortium; ANGST, Anxiety NeuroGenetics STudyConsortium; PGC, Psychiatric Genomics Consortium

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