View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Carolina Digital Repository

2019 Add Health Wave IV Documentation

Add Health

The National Longitudinal Study of Adolescent to Adult Health

Wave IV PGSs for Risk-Tolerance User Guide

Report prepared by

Richard Karlsson Linnér

Jonathan P. Beauchamp

UNC | CAROLINA POPULATION CENTER

CAROLINA POPULATION CENTER | CAROLINA SQUARE - SUITE 210 | 123 WEST FRANKLIN STREET | CHAPEL HILL, NC 27516

Add Health is supported by grant P01-HD31921 from the Eunice Kennedy Shriver National Institute of Child Health and Human Development, with cooperative funding from 23 other federal agencies and foundations.

doi.org/10.17615/m57h-bw07

Add Health is a program project directed by Kathleen Mullan Harris and designed by J. Richard Udry, Peter S. Bearman, and Kathleen Mullan Harris at the University of North Carolina at Chapel Hill, and funded by grant P01-HD31921 from the Eunice Kennedy Shriver National Institute of Child Health and Human Development, with cooperative funding from 23 other federal agencies and foundations. Information on how to obtain the Add Health data files is available on the Add Health website (http://www.cpc.unc.edu/addhealth).

SSGAC Risk Tolerance: GWAS and MTAG Polygenic Scores (Ver 1.0)

Contact: Richard Karlsson Linnér <r.karlssonlinner@vu.nl>, Jonathan P. Beauchamp <jonathan.pierre.beauchamp@gmail.com>

Date: March 19, 2019

Introduction. Karlsson Linnér *et al.*[1] conducted genome-wide association analyses of general risk tolerance (n = 975,353), adventurousness and risky behaviors in the driving, drinking, smoking and sexual domains. In separate hold-out cohorts, they analyzed the predictive power of polygenic scores derived from the genome-wide association study (GWAS) estimates. Due to data access restrictions, it is not possible to release summary statistics for more than 10,000 single nucleotide polymorphisms (SNPs). Therefore, researchers with access to the individual-level genotype data cannot reproduce the polygenic scores that were used in the paper from publicly available summary statistics (https://www.thessgac.org/data). As a partial remedy, we are releasing the polygenic scores that were used in the paper's prediction analyses in the Add Health and UKB-siblings cohorts to researchers (but due to the restrictions, we cannot release the underlying SNP-level weights themselves).

Scores for European-ancestry Add Health respondents are available here:

"KarlssonLinner_et_al_(2019)_PGS_AddHealth.txt"

The Add Health sample is described below. In that sample, Karlsson Linnér *et al.*[1] estimated the predictive power of a polygenic score based on summary statistics from the paper's primary meta-analysis of general risk tolerance (n = 975,353). If you use these scores, please cite:

Karlsson Linnér, R. *et al*. Genome-wide association analyses of risk tolerance and risky behaviors in over 1 million individuals identify hundreds of loci and shared genetic influences. *Nat. Genet.* **51**, 245–257 (2019).

The purpose of this document is to briefly describe the construction of the scores and the Add Health sample. For additional details, readers are referred to the Supplementary Note of Karlsson Linnér *et al.*[1], especially Sections 2 (GWAS, Quality Control and Meta-analysis) and 10 (Predictive Power of General-Risk-Tolerance Polygenic Score).

Methodology. A polygenic score for an individual is defined as a weighted sum of a person's genotypes at M SNPs,

$$\hat{S}_i = \sum_{j=1}^{M} g_{ij}\beta_j \qquad (1)$$

We employed two methods to generate the weights β_j. First, we used the software Plink[2] to produce classical scores. For classical polygenic scores, the estimated additive effect size $\hat{\beta}_j$ for

SNP *j* is the GWAS estimate for SNP *j*. Second, we used the LDpred[3] method, a Bayesian method that includes all measured SNPs and weights each SNP by (an approximation) of its conditional effect, given other SNPs. The theory underlying LDpred is derived assuming the variance-covariance matrix of the genotype data in the training sample is known and assuming some prior effect-size distribution. In practice, the matrix is not known but must be approximated using LD patterns from a reference sample. LDpred calculates posterior effect-size distributions for the true conditional effect sizes ***b*** (i.e., the effect sizes conditional on all other SNPs in a window), and each SNP's weight is set equal to the mean of its (conditional) posterior effect-size distribution.

Genotype data and imputation for the Add Health individuals. Genotype data from the Illumina Omni 1.1 and 2.5 chips were available for 9,975 Add Health[5] individuals and 606,673 variants. We imputed these genotypes against the Haplotype Reference Consortium (HRC) v1.1 European reference panel[4] using the Michigan Imputation Server. Prior to imputation, we identified the non-European individuals by plotting the principal components (PCs) of the covariance matrix of the individuals' genotype data together with the PCs of 1000 Genomes populations[6] and visually inspecting the plots. We dropped the identified 4,187 non-European individuals from the sample. Additionally, we excluded individuals that do not satisfy the following criteria: (i) genotype missingness rate is less than 0.05 in all chromosomes, (ii) there is no mismatch between surveyed sex and genetic sex, (iii) the individual is not an outlier in terms of heterozygosity/homozygosity, and (iv) the individual is not an ancestral outlier. We also dropped SNPs that have a call rate less than 0.98, Hardy-Weinberg exact test *P*-value less than 10^{-4}, or minor allele frequency < 0.01.

Next, we checked the data against the HRC reference panel[a] for consistency of strand, id names, positions, alleles, reference/alternative allele assignment, and allele frequency differences using version 4.2.5 of the HRC-1000G-check-bim.pl[b] program. The program updates strand, position and reference/alternative allele assignment when possible. It removes a SNP if it has any of the following properties: (i) A/T or G/C alleles and a minor allele frequency greater than 0.4, (ii) alleles that do not match the HRC data, (iii) minor allele frequency discrepancy with the HRC data greater than 0.2, (iv) not available in the HRC data. After all checks, 346,754 SNPs remained which were taken forward for imputation. Genotype probabilities were imputed for 39,117,084 variants and 5,690 individuals. To construct the scores for the individuals in our prediction sample, we used only the subset of SNPs in the HapMap consortium phase 3 release[7 8] and we used best-guess genotypes.

Polygenic scores. We provide three types of polygenic scores for risk tolerance based on different summary statistics and weight-estimation methods. All three scores were constructed from meta-analyses that did not contain the Add Health cohort:

(i) A classical score, calculated using Plink, based on standard GWAS summary statistics estimated from a univariate GWAS of general risk tolerance. That score does not take into account LD patterns in the GWAS sample. The GWAS summary statistics are from a meta-analysis that combines the discovery and replication cohorts in Karlsson Linnér *et al.*, which does not include the sample of Add Health (n = 975,353);

[a] Site list was downloaded from http://www.haplotype-reference-consortium.org/site

[b] Script available at http://www.well.ox.ac.uk/~wrayner/tools/HRC-1000G-check-bim.v4.2.5.zi

(ii) An LDpred score based on the same GWAS summary statistics as (i) (but estimated using LDpred, which accounts for LD patterns in the GWAS sample); and

(iii) An LDpred score based on the MTAG[9] summary statistics for risk tolerance, which were obtained from a multivariate analysis of the same GWAS summary statistics of risk tolerance as (i) (*n* = 975,353)*;* adventurousness (*n* = 557,923); automobile speeding propensity (*n* = 404,291); drinks per week (*n* = 414,343); ever smoker (*n* = 518,633); number of sexual partners (*n* = 370,711); and lifetime cannabis use (*n* = 32,330) (see Table 1 below and Supplementary Information Section 10 of Karlsson Linnér *et al.*[1] for additional details).

For the two LDpred scores, we adjusted the weights for linkage disequilibrium using the LDpred software tool[3] and the reference genotype data whose construction is described below. The LD-adjusted univariate GWAS weights were obtained for SNPs that are available in both the reference data and the standard GWAS summary statistics for the phenotype, and that pass the filters imposed by LDpred: (i) the variant has a minor allele frequency (MAF) greater than 1% in the reference data, (ii) the variant does not have ambiguous nucleotides, (iii) there is no mismatch between nucleotides in the summary statistics and reference data, and (iv) there is no high (>0.15) MAF discrepancy between summary statistics and validation sample. The LD-adjusted MTAG weights were further restricted to SNPs that are available in the GWAS summary statistics for all seven phenotypes. The number of SNPs included in the classical score and LDpred GWAS score is 1,167,185. The number of SNPs included in the LDpred MTAG score is 1,110,220. The posterior effect sizes were calculated assuming a fraction of causal SNPs equal to 0.3 and setting the LD window to *M*/3000, where *M* is the number of SNPs included in the score.

We completed the last step of calculating the scores in Plink v1.9[2], using the *Add Health* individuals' best-guess genotype data and the LD-adjusted weights described above, for 4,755 *Add Health* individuals.

<u>*Estimation of LD patterns (for the two LDpred scores).*</u> We estimated LD patterns using the HRC (Haplotype Reference Consortium) Genomes-imputed genotype data (Version 1.1) of *Add Health*. To obtain the LD reference data, we converted the genotype probabilities for 38,898,725 biallelic SNPs to hard calls using Plink v1.9[2]. We restricted the set of genetic variants to 1,211,662 HapMap3[7] SNPs, because these SNPs are generally well-imputed and provide good coverage of the genome in European-ancestry individuals. Next, we estimated a genetic relatedness matrix, restricting further to SNPs with minor allele frequency greater than 0.01. We dropped one individual from each of the 934 pairs of individuals with a genetic relatedness exceeding 0.02.

In order to make sure that there are no genetic outliers in the sample that can bias the LD estimates, we clustered the remaining 4,756 individuals based on identity-by-state distances in Plink v1.9[2], again restricting to SNPs with minor allele frequency greater than 0.01. Plink reports a *Z*-score for each individual's identity-by-state distance to his/her closest neighbor. We examined these *Z*-scores and marked an individual as genetic outlier if his/her *Z*-score was smaller than -5. One such individual was identified who was then dropped from the sample. The process was repeated, confirming that no individual with a *Z*-score less than -5 remained in the sample. In the final data set, there were 4,755 individuals and 1,211,662 SNPs.

Table 1. Phenotype measures

Phenotype	Measure	Uses
General risk tolerance	Meta-analysis of UK Biobank, 23andMe and 10 smaller cohorts: UK Biobank[10]: Would you describe yourself as someone who takes risks? [1] Yes, [2] No 23andMe[11]: In general, people often face risks when making financial, career, or other life decisions. Overall, do you feel comfortable or uncomfortable taking risks? [1] Very comfortable, [2] Somewhat comfortable, [3] Neither comfortable nor uncomfortable, [4] Somewhat uncomfortable, [5] Very uncomfortable Detailed measures of general risk tolerance for the 10 smaller cohorts are in the Supplementary Table 4 of Karlsson Linnér et al.[1]	GWAS of general risk tolerance + MTAG analysis with general risk tolerance as the primary phenotype
Adventurousness	If forced to choose, would you consider yourself to be more cautious or more adventurous? [1] Very cautious, [2] Somewhat cautious, [3] Neither, [4] Somewhat adventurous, [5] Very adventurous	MTAG analysis with general risk tolerance as the primary phenotype
Automobile speeding propensity	How often do you drive faster than the speed limit on the motorway? [1] Never/rarely, [2] Sometimes, [3] Often, [4] Most of the time, [5] Do not drive on the motorway We first dropped all participants who reported not driving on the motorway, and then we normalized our categorical variable for males and females separately.	MTAG analysis with general risk tolerance as the primary phenotype
Drinks per week	Our drinks per week measure is constructed from responses to a sequence of questions in the UK Biobank[10]. First, respondents were asked how often they drink alcohol, and response options include [1] Daily or almost daily, [2] Three or four times per week, [3] Once or twice per week, [4] One to three times per month, [5] Special occasions only, and [6] Never Respondents who reported drinking once per week or more were asked how many glasses of various types of alcoholic beverages they consume per week. We used the sum of all alcoholic drinks per week as our drinks per week phenotype for these respondents. Respondents who reported drinking less than once per week	MTAG analysis with general risk tolerance as the primary phenotype

	(one to three times per month or on special occasions only) were asked how many glasses of various types of alcoholic beverages they consume per month. For these respondents, we added the total number of drinks per month and divided by 4 to arrive at an approximated number of drinks per week. Respondents who reported never drinking were coded as 0.	
Ever smoker	Meta-analysis of the following two studies: UK Biobank[10]: we coded ever-tobacco smoker status as 1 if a respondent reported that they were a current or previous smoker and 0 if they reported never smoking or only smoking once or twice. We coded cigarettes per day as 0 if ever-smoking status was also 0; otherwise, we used the maximum number of reported past or current cigarettes (or pipes/cigars) consumed per day, normalized separately for males and females. Tobacco, Alcohol and Genetics (TAG) Consortium[12]: A published meta-analysis of 16 cohorts.	MTAG analysis with general risk tolerance as the primary phenotype
Number of sexual partners	About how many sexual partners have you had in your lifetime? If respondents reported more than 99 lifetime sexual partners, they were asked to confirm their responses. We assigned a value of 0 to participants who reported having never had sex, and we again normalized this measure separately for males and females.	MTAG analysis with general risk tolerance as the primary phenotype
Lifetime cannabis use	GWAS summary statistics from a published study.[13]	MTAG analysis with general risk tolerance as the primary phenotype

A note on the MTAG-based polygenic score. MTAG[9] is a method that uses GWAS summary statistics for a primary phenotype and for one or more secondary phenotypes to produce an updated set of summary statistics for the primary phenotype which, under certain assumptions, will be more precisely estimated than the input GWAS summary statistics.

There are costs and benefits to using an MTAG-based polygenic score. For instance, in all cases, MTAG-based polygenic scores will be more predictive of their corresponding phenotype in expectation. In some cases, however, MTAG can have a high false discovery rate (see

Supplementary Note section 1.4 of Turley *et al.*[9]), which may lead to spurious correlations between the MTAG-based polygenic score and other phenotypes.

We therefore offer the following recommendations. If in a regression, the dependent variable and the polygenic score correspond to the same phenotype, we recommend using the MTAG-based score. If the dependent variable and the polygenic score correspond to different phenotypes, but the coefficient of interest in the regression is not the coefficient associated with the polygenic score (e.g., if the polygenic score is only being used as a control variable in an experimental setting), then we also recommend using the MTAG-based polygenic score. Care should be taken when interpreting the coefficient of an MTAG-based polygenic score in this setting, however, since any observed association may be driven through channels involving the secondary phenotypes. This is especially true when the maxFDR is large (see Turley *et al.*[9], Supplementary Note section 1.4). If researchers are interested in the coefficient on the polygenic score, they should either use GWAS-based scores, or justify why such channels would lead to negligible bias in their particular case.

Principal components. It is important to take a number of steps to minimize the risk that an observed association between the outcome of interest and the polygenic score is due to unaccounted-for population stratification. A score is stratified if its distribution varies across members of different ancestry groups. Failure to control for differences in ancestry can severely bias estimates of effect sizes, since members of different groups may vary in the outcome of interest for environmental reasons[14]. To reduce such concerns, we recommend controlling for the top 10 ancestry-specific principal components (PCs) of the covariance matrix of the individuals' genotypic data[15], which are included in "KarlssonLinner_et_al_(2019)_PGS_AddHealth.txt".

Variables. Table 2 provides a description of the variables included in "KarlssonLinner_et_al_(2019)_PGS_AddHealth.txt".

Table 2. Description of variables

Variable	**Description**
aid	Individual identifier
FID	Family identifier
PGS_RISK_PLINK_GWAS	Polygenic score for general risk tolerance, obtained using classic PLINK method and standard GWAS results
PGS_RISK_LDPRED_GWAS	Polygenic score for general risk tolerance, obtained using LDpred method and standard GWAS results
PGS_RISK_LDPRED_MTAG	Polygenic score for general risk tolerance, obtained using LDpred method and results from multivariate analysis of adventurousness, automobile speeding propensity, drinks per week, ever smoker, number of sexual partners, and lifetime cannabis use

PC1 - PC10	Top 10 principal components (PCs) of the covariance matrix of the individuals’ genotypic data

References

1. Karlsson Linnér, R. *et al.* Genome-wide association analyses of risk tolerance and risky behaviors in over 1 million individuals identify hundreds of loci and shared genetic influences. *Nat. Genet* **51**, 245-257. (2019).

2. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4,** 7 (2015).

3. Vilhjálmsson, B. J. *et al.* Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. *Am. J. Hum. Genet.* **97,** 576–592 (2015).

4. The Haplotype Reference Consortium. A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48,** 1279-1293 (2016).

5. Harris, Kathleen Mullan, Carolyn Tucker Halpern, Eric A. Whitsel, Jon M. Hussey, Ley Killeya-Jones, Joyce Tabor, and Sarah C. Dean. 2019. "Cohort Profile: The National Longitudinal Study of Adolescent to Adult Health (Add Health)." *International Journal of Epidemiology* (published online June 29, 2019) https://doi.org/10.1093/ije/dyz115.

6. A global reference for human genetic variation, The 1000 Genomes Project Consortium, Nature **526**, 68-74 (01 October 2015) doi:10.1038/nature15393.

7. Altshuler, D. M., Gibbs, R. A. & Peltonen, L. Integrating common and rare genetic variation in diverse human populations. *Nature* **467,** 52–58 (2010).

8. Buchanan, C. C., E. S. Torstenson, W. S. Bush, and M. D. Ritchie. A comparison of cataloged variation between International HapMap Consortium and 1000 Genomes Project data. *Journal of American Medical Informatics Association* 19, 289-294 (2012).

9. Turley, P. *et al.* Multi-trait analysis of genome-wide association summary statistics using MTAG. *Nat. Genet.* **50,** 229-237 (2018). doi:10.1038/s41588-017-0009-4

10. Bycroft, C. *et al.* Genome-wide genetic data on ~500,000 UK Biobank participants. *bioRxiv* 166298 (2017). doi:10.1101/166298

11. Hinds, D. *et al*. Genetic Discovery in the 23andMe Participant Cohort. 23andMe, Inc. (2014).

12. The Tobacco and Genetics Consortium. Genome-wide meta-analyses identify multiple loci 161 associated with smoking behavior. *Nat. Genet.* **42**, 441–447 (2010).

13. Stringer, S. *et al*. Genome-wide association study of lifetime cannabis use based on a large meta-analytic sample of 32,330 subjects from the International Cannabis Consortium. *Transl. Psychiatry* **6**, e769 (2016).

14. Hamer, D. & Sirota, L. Beware the chopsticks gene. *Mol. Psychiatry* **5,** 11–13 (2000).

15. Price, A. L. *et al.* The impact of divergence time on the nature of population structure: an example from Iceland. *PLoS Genet.* **5,** e1000505 (2009).