MIDUS
Midlife in the United States
A National Study of Health & Well-Being

MIDUS Project 6 (Genetics):

Polygenic Risk Score (PRS) Documentation

For File:

M2MR_P6_PolygenicRiskScores_N2118_20190603

October 2019

NIA Grant #U19 AG051426 and post-doctoral support from a Templeton Foundation Grant.

X:\Administrative Database\DBS\MIDUS 3\Project 6 (Genetics)\DNA\Polygenic Scores\Documentation\M2MR_P6_PolygenicRiskScores_Documentation_20191004.docx

Polygenic Risk Score (PRS) Data

Overview

DNA was extracted from tissue samples (whole blood, saliva) obtained from MIDUS 2 (M2) and MIDUS Refresher (MR) Biomarker participants and then genotyped using the Illumina Omni Express array. The resultant SNPs were then processed via established protocols for genotype calling[A], ancestry measurement, and imputation of genotypes and a set of Polygenic Risk Scores (PRS) was computed. The purpose of this document is to provide information about the PRS, how they were generated, and considerations that must be made when using PRS in analyses. General information about the data file is provided first, and subsequent sections provide additional details about sample collection and processing as well as computation of the PRS.

Sample and Variables

In general, MIDUS data are processed and distributed separately by wave of data collection (M1, M2, Refresher etc.). At the time that DNA extraction was initiated, tissue samples were available from both M2 and MR. SNP data from both waves were combined for efficiency of processing, so the PRS data file described here contains data for M2 and MR cases. Consistent with other MIDUS files, the file includes the full set of M2 (n=1255) and MR (n=863) biomarker cases for a total sample size of N=2118. The file includes a flag variable (see below) so that cases with PRS data may be easily identified.

MIDUS variable naming and coding conventions (seen Naming and Coding Conventions included with the Survey documentation) specify that the first 3-4 characters in a variable name indicate the wave (i.e. B for M2, RA for Refresher wave 1), the project, and the instrument or data type. Since the PRS data file includes data from both M2 and MR, the convention is modified slightly such that each variable begins with BRA. The data were generated under the auspices of the Genetics project (P6) and are derived from DNA, thus the first 5 characters of the variables, except for MIDUS Administrative variables, begin with BRA6D. The remaining characters are determined by the data. For example, variables derived from principal components analysis include 'PC' while the PRS variables include 'PRS'.

The SPSS data file "M2MR_P6_PolygenicRiskScores_N2118_20190603.sav" contains 64 variables as follows:

- Administrative: there are 6 variables
    - M2MRID – contains the public identifiers for the MIDUS core sample and the MIDUS Refresher
    - SAMPLMAJ – the identifier created by the MIDUS Administrative Core to indicate the participants 'sample of origin' (e.g. MIDUS Refresher, Twin, etc.),
    - M2MRCASE – indicates if sample was obtained as part of the MIDUS 2 or MIDUS Refresher data collection
    - GENCONSENT – flag variable indicating whether the participant consented to genetics or not.

[A] "In next-generation sequencing (NGS) methods, a whole genome, or targeted regions of the genome, is randomly digested into small fragments (or short reads) that get sequenced and are then either aligned to a reference genome or assembled. Having aligned the fragments of one or more individuals to a reference genome, 'SNP calling' identifies variables sites, whereas 'genotype calling' determines the genotype for each individual at each site." (p. 443, Nielsen, Paul, Albrechtsen, & Song, 2011)

    - BRA6DTISSUE – indicates the tissue sample source from which DNA was extracted. See below for information about sample collection.
    - BRA6DPAVAIL – flag variable indicating if PRS are available for a given case or not.
- Principal Component Scores (PC): Genotype principal component analysis was performed as part of determination of ancestry. The top principal components (PCs) from that process (see details below) are included in the current data file (BRA6DPC1 to BRA6DPC5).
    - All 5 of these PC variables should be included in PRS analyses, along with age and sex. For an introduction to the interpretation of PC scores, see Novembre & Stephens (2008).
    - Ancestry Scores: there are 5 ancestry component scores (BRA6DPEUR -European, BRA6DPCEAS - East Asian, etc.). Ancestry component scores range from 0 to 1 and represent the probability of one's ancestry based on their genotype. See below for computation details.
- Polygenic Risk Scores (PRS) – there are 47 PRS variables. The remaining characters in each variable name serve as a unique identifier for the phenotype (e.g. BRA6DPRSALZ "ALZ" for Alzheimer's Disease, BRA6DPRSAOM "AOM" for Age of Menarche, etc.)
    - PRS are only included for individuals of predominately European descent, due to ancestry confounds present in non-European data. See details about PRS computation below.
    - Table 1 below contains a list of phenotypes for which PRS data are available and associated references for discovery genome-wide association studies.

DNA Tissue Samples

Collection and storage of tissue samples for genetics was added to the MIDUS Biomarker protocol about halfway through the data collection period. Once that change was implemented, whole blood was stored for DNA extraction for the subset of MIDUS 2 participants and all the MIDUS Refresher participants who consented to genetic testing at the clinic visit. About half of the M2 participants completed the clinic visit before genetics was added to the protocol. At the end of the M2 data collection period individuals in this group were invited to provide saliva samples for genetic testing by mail using Oragene DNA Self-Collection Kits (OG-300). Saliva samples were obtained from about 70% of that group.

Analytic Guidelines

Background: PRS provide an index of an individual's genetic propensity for a given phenotype based on common genetic variation. Put differently, PRSs provide "an individual-level genome-wide genetic proxy"[1] for a variable of interest. PRSs can be computed for any variable for which there has been a large discovery genome wide association study (e.g. height, weight, BMI, LDL, HDL, triglycerides, type-II diabetes, schizophrenia, subjective well-being, etc.) and can be used in a variety of applications. Summary data used to generate PRSs can be used to examine shared genetic covariation among two or more variables, epistasis or genome x genome interactions between more or more variables, and genome x environment interactions.[1,2] PRS summary data are also commonly used as instrumental variables to infer quasi-causal relations between exposures and outcomes in observational studies, sometimes called Mendelian randomization studies[3,4]. In all of these applications, the calculation of a polygenic risk score depends on estimates from a genome-wide association study (GWAS), specifically the effect sizes of single nucleotide polymorphisms (SNPs; β_i) that are statistically associated with the phenotype of interest. As a result, the predictive power and validity of polygenic risk score analyses depend on both the size and quality of the discovery GWAS and the degree those samples match the target sample in which the PRS analyses are performed.

IMPORTANT: PRS analyses are highly prone to ancestry bias, and the discovery GWAS from which SNP weights are obtained are almost universally based on participants of Non-Finnish, Northern European ancestry. Until non-European ancestry GWAS summary statistics are both available and reliable, PRS analyses should only be conducted with samples of European participants. **In addition, all PRS analysis requires inclusion of ancestry principal components as covariates, along with other critical covariates (age, sex, etc).**

When conducting PRS analyses in MIDUS, it is also important to account for the nested structure of the data, as genotyped participants include twins (~ 100 pairs) that were raised in the same home. PRS analyses of twin data should typically only be considered in studies comparing sibling pairs. In other cases, one twin should be left out of analyses. PRS will be highly or completely overlapping across siblings due to the large amount of DNA shared and the way in which PRS aggregate common variation.

When conducting GxE and GxG studies, it is important to make sure that potential confounding variables are accounted for. "To properly control for confounders, researchers need to enter the covariate-by-environment and the covariate-by-gene interaction terms in the same model that tests the G×E term." (Keller, 2014). For example, if a researcher wanted to test a hypothesis about a particular polygenic risk score (PRS) interacting with a measured environment (E), and wanted to control for three potential confounders (cov1 - cov3), then the following regression equation should be estimated:

$$y = b_o + b_1PRS + b_2E + b_3cov1 + b_4cov2 + b_5cov3 + b_6PRS\times E + b_7cov1\times PRS + b_8cov2\times PRS + b_9cov3\times PRS + b_{10}cov1\times E + b_{11}cov2\times E + b_{12}cov3\times E + e$$

b_0 = intercept
b_1-b_{12} = slope
e = residual error term

The same procedure should be followed to control for potential confounders when testing G×G interactions. When testing whether a particular variable statistically mediates or accounts for polygenic liability for a phenotype (i.e. when searching for endophenotypes), **it is important to bootstrap standard errors when estimating indirect effects**. This can be easily achieved using modern statistical software, including SPSS, SAS, R, and Mplus. Finally, when testing multiple PRS → phenotype associations, **it is important to adjust p-value thresholds for multiple comparisons** using a Bonferroni correction ($p < .05$/total number of comparisons). In some cases, FDR correction may be used. Technical details about steps for generating the PRSs are provided in the following section.

Generating Polygenic Risk Scores

Genotyping Quality Control: Genotyping quality control (QC) for MIDUS samples was performed on approximately 500k common variants using single nucleotide polymorphism (SNP) clustering in the Illumina Genome Studio software[B]. Samples were assayed on 2 different versions of an Illumina OmniExpress array (1.0 and 1.1) and, consequently, SNP calling using Genome Studio required separate processing from different arrays. The data from the different arrays had previously been processed jointly, resulting in major bias and a loss of quality. To correct this, each batch was re-processed separately with Genome Studio, using the appropriate clustering files for each respective genotyping array. This significantly increased SNP call rate and improved other QC metrics.

[B] https://www.illumina.com/techniques/microarrays/array-data-analysis-experimentaldesign/genomestudio.html

Ancestry Analysis: Ancestry of the MIDUS participants was estimated using Admixture software[C] with a 1000 Genomes data (Phase 3) reference.[5] All 5 super-populations[D] were used as a basis for estimation. To calculate ancestry component scores, genotype principal components analysis (PCA) was performed on MIDUS genotypes after linkage disequilibrium (LD) pruning SNPs at a 0.2 R^2 threshold. The PCA was run using RaMWAS[6], a Bioconductor[7] package which comprises a complete toolset for QC, GWAS and methylome-wide association studies. RaMWAS includes functions for PCA for capturing batch effects and detection of outliers, association analysis while correcting for top PCs and covariates, creation of QQ-plots and Manhattan plots, and annotation of significant results. In total, five ancestry component scores were calculated: European (EUR), East Asian (EAS), Ad-mixed American (AMR), Southeast Asian (SAS), and African (AFR).

Inclusion Criteria: To date, discovery GWASs have focused almost exclusively on participants of Non-Finnish Northern European ancestry. Consequently, the estimated effect sizes of individual SNPs are only known for individuals of European ancestry, and, as a result, the calculation of polygenic scores are only valid for participants of predominantly European ancestry. Therefore, to exclude ancestrally heterogeneous samples from the data, the top principal components (PCs), defined as those components which accounted for > 0.1% of the genotype variance, $n_{pc} = 4$, were used to establish PC centroid limits centered around 1000 Genomes CEU data such that 99% of the CEU data fell within the limits. Only MIDUS samples also falling within these limits (N = 1309) were considered ancestrally homogenous and, thus, were included in polygenic risk scoring.

The Illumina OmniExpress arrays tag a sufficient number of variants on the X and Y chromosomes to determine biological sex (e.g. 17,707 SNPs on X chromosome and 1,367 on Y for array v. 1.1). Samples were excluded if self-reported sex did not match biological sex as determined by genotype (N = 13), as this indicates either invalid self-reports, genotyping errors, or accidental I.D. swaps. In sum, after filtering out samples that did not pass ancestry- and sex-checks, PRSs were calculated for a final sample of N = 1296 participants.

Genotype Imputation: After MIDUS genotype samples were filtered via inclusion criteria, genotypes were imputed to approximately 80 million SNPs using minimac3[8] and Eagle[9] with the 1000 Genomes reference panel[5] on the Michigan Imputation Server.[8] SNPs with ambiguous strand orientation, >5% missing calls, or Hardy-Weinberg equilibrium $p < 0.001$ were excluded prior to imputation. After imputation, SNPs with minor allele frequency below 0.01 or an average call rate (AvgCall) below 0.9 were excluded, resulting in ~8 million common variants with sufficient minor allele frequency and call rate for inclusion. This variant calling pipeline follows the established best practice methods from the Broad Institute, incorporated in the freely available Genome Analysis Toolkit (**GATK**)[10]. All genomic data were handled using Plink 1.9[11,12].

Polygenic Risk Scoring:
To be more precise, a PRS may be defined "as a single value estimate of an individual's propensity to a phenotype, calculated by computing the sum of risk alleles corresponding to a phenotype of interest in each individual, weighted by the effect size estimate of the most powerful GWAS on the phenotype" [1]. PRS is traditionally calculated as $\mathrm{PRS}_k = \sum_i \beta_i \, \mathrm{SNP}_{ik}$, where PRS for individual *k* in the target sample is

[C] http://software.genetics.ucla.edu/admixture/admixture-manual.pdf

[D] 26 different populations from around the world comprise the 1000 Genomes reference data, and these populations are divided into 5 "super-populations": African, Ad Mixed American, East Asian, European, and South Asian. For a list of populations and their descriptions, see: http://www.internationalgenome.org/category/population/

calculated by the summation of each SNP (measured for both the person *k* and passing a set association threshold in the discovery GWAS) multiplied by the effect size, *β*, of that SNP in the discovery GWAS. PRSs are then typically transformed to a standardized metric (M = 0, SD = 1).

PRSice 2.0[E,13] was used to calculate polygenic risk scores. A default threshold of p = 1.0 (including all SNPs of infinitesimal effect) was used to calculate scores. A variety of thresholds can be used, and some researchers prefer a threshold of p = 0.5. Thresholds are not arbitrary. We adopt a more liberal threshold of p = 1.0, due to the diverse array of medical and psychiatric phenotypes, each with varying, undefined levels of genetic complexity and a tendency toward massive polygenicity. Stricter thresholds (0.3, 0.5) may be needed for any given PRS, particularly if there is precedence in the literature for such a prediction. If multiple thresholds for PRS are used, analytic p-values must be corrected for multiple comparison. Approximately 4000 phenotypes are currently available for PRS calculation, but for the MIDUS samples, only phenotypes with adequate discovery GWAS sample size (*N* ~ 10,000) were selected for polygenic risk scoring. This helps to minimize the standard error of the SNP weights used in the calculation of PRSs and increases predictive power. See Table 1 below for a list of phenotypes for which PRS data is available and associated references for discovery genome-wide association studies. Note, because this field is changing rapidly, researchers should use these references as a starting point and regularly check the literature for more recent advances.

[E] https://choishingwan.github.io/PRSice/

Table. Phenotypes and Samples Sizes from Discovery GWAS

Phenotype	N cases	N controls	N total
ADHD Females[14]	30,961	938,949	969,910
ADHD Males[14]	68,694	969,312	1,038,006
Adventurousness[15]	-	-	557,923
Age at Menarche[16]	-	-	370,000
Alzheimer's Disease[17]	25,580	48,466	74,046
Anorexia Nervosa[18]	3,495	10,982	14,477
Anxiety[19]	-	-	18,186
Asthma[20]	10,365	16,110	26,475
Autism Spectrum Disorder[21]	18,381	27,969	46,350
Automobile Speeding Propensity[15]	-	-	404,291
Bipolar Disorder[22]	29,764	169,118	198,882
Birth Height[23]	-	-	28,495
Birth Weight[24]	-	-	153,781
Body Mass Index[25]	-	-	234,069
Child IQ[26]	-	-	17,989
Cigarettes Per Day[27]	-	-	73,853
Cognitive Performance[28]	-	-	257,841
College Education[29]	20,040	75,387	95,427
Coronary Artery Disease[30]	22,233	64,762	86,995
Crohn's Disease[31]	6,299	15,148	21,447
Cross-Disorder[32]	33,332	27,888	61,220
Depressive Symptoms[33]	-	-	180,866
Disinhibition[15]	-	-	315,894
Drinks Per Week[15]	-	-	414,343
Educational Attainment[28]	-	-	766,344
Ever Smoked[15]	-	-	518,633
Extraversion[34]	-	-	63,030
General Risk Tolerance[15]	-	-	557,923
HDL Cholesterol[35]	-	-	99,900
Height[36]	-	-	253,288
Infant Head Circumference[37]	-	-	10,768
Intracranial Volume[38]	-	-	30,717
LDL Cholesterol[35]	-	-	95,454
Major Depressive Disorder[39]	59,851	113,154	173,005
Neuroticism[40]	-	-	168,105
Number of Sexual Partners[15]	-	-	370,711
Obsessive-Compulsive Disorder[41]	2,688	7,037	9,725
Post-Traumatic Stress Disorder[42]	2,488	7,466	9,954
Rheumatoid Arthritis[43]	5,539	20,169	25,708
Risk Tolerance[15]	-	-	939,908
Schizophrenia[44]	36,989	113,075	150,064

Subjective Well-Being[33]	-	-	298,420
Suicide Death*			
Total Cholesterol[34]	-	-	100,184
Triglycerides[34]	-	-	96,598
Type 2 Diabetes[45]	62,892	596,424	659,316
Ulcerative Colitis[46]	7,211	20,783	27,994
Waist-to-Hip Ratio[47]	-	-	142,762

*Contact lab for archival citation

References

1. Choi, S. W., Mak, T. S. H., & O'reilly, P. (2018). A guide to performing Polygenic Risk Score analyses. *BioRxiv*, 416545.
2. Dudbridge, F. (2013). Power and predictive accuracy of polygenic risk scores. *PLoS genetics*, *9*(3), e1003348.
3. Davey Smith, G., & Hemani, G. (2014). Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Human molecular genetics*, *23*(R1), R89-R98.
4. Bowden, J., Davey Smith, G., & Burgess, S. (2015). Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *International journal of epidemiology*, *44*(2), 512-525.
5. The Genomes Project Consortium. A global reference for human genetic variation. *Nature* 526, 68 (2015).
6. Shabalin, A., Clark, S., Hattab, M., Aberg, K. & van den Oord, E. *RaMWAS: Fast Methylome-Wide Association Study Pipeline for Enrichment Platforms* . R package version 1.2.0 edn (2017).
7. Bioconductor. Bioconductor: Open Source Software for Bioinformatics. Vol. 2017 (2017).
8. Das, S. *et al.* Next-generation genotype imputation service and methods. *Nat Genet* 48, 1284-1287 (2016).
9. Loh, P.R. *et al.* Reference-based phasing using the Haplotype Reference Consortium panel. *Nat Genet* 48, 1443-1448 (2016).
10. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 20, 1297-303 (2010).
11. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81, 559-75 (2007).
12. Purcell, S., Chang, C., NIH-NIDDK Laboratory of Biological Modeling & Purcell Lab at Mount Sinai School of Medicine. PLINK 1.9 beta. Vol. 2017 (2017).
13. Euesden, J., Lewis, C.M. & O'Reilly, P.F. PRSice: polygenic risk score software. *Bioinformatics* 31, 1466-1468 (2014)
14. Martin, J. *et al.* A genetic investigation of sex bias in the prevalence of attention-deficit/hyperactivity disorder. *Biological psychiatry* 83, 1044-1053 (2018).
15. Linnér, R.K. *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. *Nature genetics*, 1 (2019).
16. Day, F.R. *et al.* Genomic analyses identify hundreds of variants associated with age at menarche and support a role for puberty timing in cancer risk. *Nature genetics* 49, 834 (2017).
17. Lambert, J.-C. *et al.* Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nature genetics* 45, 1452 (2013).
18. Duncan, L. *et al.* Significant locus and metabolic genetic correlations revealed in genome-wide association study of anorexia nervosa. *American journal of psychiatry* 174, 850-858 (2017).

19. Otowa, T. *et al.* Meta-analysis of genome-wide association studies of anxiety disorders. *Molecular psychiatry* 21, 1391 (2016).
20. Moffatt, M.F. *et al.* A large-scale, consortium-based genomewide association study of asthma. *New England Journal of Medicine* 363, 1211-1221 (2010).
21. Grove, J. *et al.* Common risk variants identified in autism spectrum disorder. *bioRxiv*, 224774 (2017).
22. Stahl, E. *et al.* Genomewide association study identifies 30 loci associated with bipolar disorder. *bioRxiv*, 173062 (2018).
23. van der Valk, R.J. *et al.* A novel common variant in DCST2 is associated with length in early life and height in adulthood. *Human molecular genetics* 24, 1155-1168 (2014).
24. Horikoshi, M. *et al.* Genome-wide associations for birth weight and correlations with adult disease. *Nature* 538, 248 (2016).
25. Locke, A.E. *et al.* Genetic studies of body mass index yield new insights for obesity biology. *Nature* 518, 197-206 (2015).
26. Benyamin, B. *et al.* Childhood intelligence is heritable, highly polygenic and associated with FN6BP1L. *Molecular psychiatry* 19, 253 (2014).
27. Tobacco and Genetics Consortium. Genome-wide meta-analyses identify multiple loci associated with smoking behavior. *Nat Genet* 42, 441-7 (2010).
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. *Nature genetics* 50, 1112 (2018).
29. Rietveld, C.A. *et al.* GWAS of 126,559 individuals identifies genetic variants associated with educational attainment. *science* 340, 1467-1471 (2013).
30. Schunkert, H. *et al.* Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nature genetics* 43, 333 (2011).
31. Jostins, L. *et al.* Host–microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* 491, 119 (2012).
32. Consortium, C.-D.G.o.t.P.G. Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. *The Lancet* 381, 1371-1379 (2013).
33. Okbay, A. *et al.* Genetic variants associated with subjective well-being, depressive symptoms, and neuroticism identified through genome-wide analyses. *Nature genetics* 48, 624 (2016).
34. 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* 46, 170 82 (2016).
35.Teslovich, T.M. *et al.* Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* 466, 707-13 (2010).
36. Wood, A.R. *et al.* Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat Genet* 46, 1173-86 (2014).
37. Taal, H.R. *et al.* Common variants at 12q15 and 12q24 are associated with infant head circumference. *Nat Genet* 44, 532-538 (2012).
38. Hibar, D.P. *et al.* Common genetic variants influence human subcortical brain structures. *Nature* 520, 224-229 (2015).
39. Wray, N.R. *et al.* Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nature genetics* 50, 668 (2018).

40. Turley, P. *et al.* Multi-trait analysis of genome-wide association summary statistics using MTAG. *Nature genetics* 50, 229 (2018).
41. Genetics, I.O.C.D.F. *et al.* Revealing the complex genetic architecture of obsessive compulsive disorder using meta-analysis. *Molecular psychiatry* 23, 1181 (2018).
42. Duncan, L.E. *et al.* Largest GWAS of PTSD (N= 20 070) yields genetic overlap with schizophrenia and sex differences in heritability. *Molecular psychiatry* 23, 666 (2018).
43. Stahl, E.A. *et al.* Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. *Nat Genet* 42, 508-14 (2010).
44. Ripke, S. *et al.* Biological insights from 108 schizophrenia-associated genetic loci. *Nature* 511, 421 (2014).
45. Xue, A. *et al.* Genome-wide association analyses identify 143 risk variants and putative regulatory mechanisms for type 2 diabetes. *Nature communications* 9, 2941 (2018).
46. Jostins, L. *et al.* Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* 491, 119-24 (2012).
47. Shungin, D. *et al.* New genetic loci link adipose and insulin biology to body fat distribution. *Nature* 518, 187-196 (2015).