Exploring Environmental and Genetic Influences on Substance Abuse Behavior

Damani Johnson, B.S.*, Ryan Gutierrez, Ph.D., Danielle Boisvert, Ph. D., Rachel Houston, Ph.D
Sam Houston State University

Sam Houston State University™ MEMBER THE TEXAS STATE UNIVERSITY SYSTEM
Department of Criminal Justice & Criminology
College of Criminal Justice

Background

The environmental and genetic contributions to substance abuse and dependence are well documented, often determined using twin or family studies. While there have been biological studies to identify genes that may influence this behavior, the majority have focused on individual genes or loci, and the cumulative impact of the combination of these has not been as thoroughly researched. More widely used in medical research to understand genetic diseases, the Genome-Wide Association Study (GWAS) is a tool used to investigate these correlations and calculate the relative weight of each gene's correlation. A GWAS typically involves a sample size of several thousands or hundreds of thousands of people to achieve statistically significant associations. Often GWAS data are publicly available and can be used as a basis for the development of a Polygenic Risk Score (PRS). A PRS attempts to describe the combined influences of a variety of genetic markers on an observed trait or behavior. This can be performed with a considerably smaller sample group, using weights determined by the GWAS to model and predict the occurrence of the trait.

Behavior Survey

Behavioral data for this study was acquired as part of a larger survey conducted by the SHSU College of Criminal Justice. 872 student participants answered more than 900 questions assessing a wide range of antisocial or harmful behaviors. 20 of the questions inquired about drug or alcohol use, abuse, and dependence. 10 of these were coded as binary yes (1) or no (0), while the remaining were quantitative answers coded 0-8 depending on frequency of use of a particular substance.

References

Agrawal, A., Neale, M., Prescott, C., Kendler, K. (2004). A twin study of early cannabis use and subsequent use and abuse/dependence of other illicit drugs. *Psychological Medicine, 34*(7), 1227-1237.

Chang, C. C., Chow, C.C., Tellier, L.C.A.M., Vattikuti, S, Purcell, S.M., Lee, J.J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience*, 4.

Choi, S.W., Mak, T.SH. & O'Reilly, P.F. Tutorial: a guide to performing polygenic risk score analyses (2020). Nat Protoc., 15, 2759–2772.

Marees, A.T., de Kluiver, H., Stringer, S., et al. A tutorial on conducting genome-wide association studies: Quality control and statistical analysis (2018). Int J Methods Psychiatr Res., 27:e1608.

Sanchez-Roige, S., Palmer, A. A., Fontanillas, P., Elson, S. L., 23andMe Research Team, the Substance Use Disorder Working Group of the Psychiatric Genomics Consortium, Adams, M. J., Howard, D. M., Edenberg, H. J., Davies, G., Crist, R. C., Deary, I. J., McIntosh, A. M., & Clarke, T. K. (2019). Genome-Wide Association Study Meta-Analysis of the Alcohol Use Disorders Identification Test (AUDIT) in Two Population-Based Cohorts. *American journal of psychiatry*. 176(2), 107–118.

Vink, J.M., Hottenga, J.J., de Geus, E.J.C., Willemsen, G., Neale, M.C., Furberg, H. and Boomsma, D.I., (2014). Polygenic risk scores for smoking. *Addiction,* 109: 1141-1151.

Results

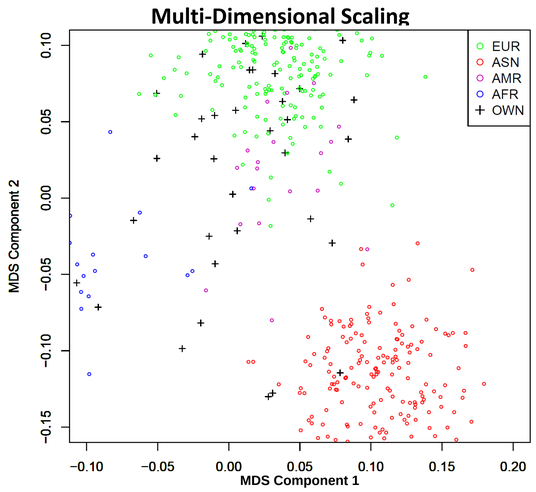

Figure 1: 1000 Genome Project data were combined with target data of all ancestries for Multi-Dimensional Analysis of population stratification.

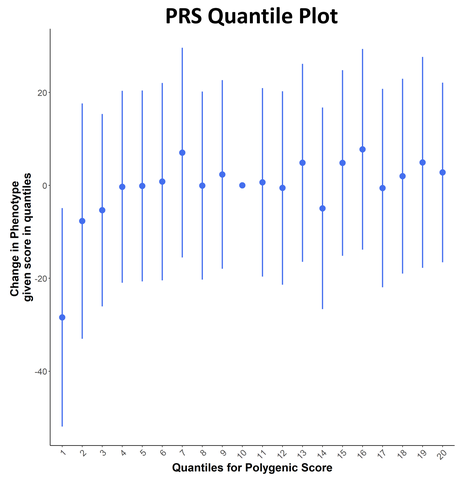

Figure 2: The sample data was separated into quantiles and plotted to observe the trend between the calculated risk scores and the dependence on alcohol. This model used only Caucasian samples and alcohol-based questions.

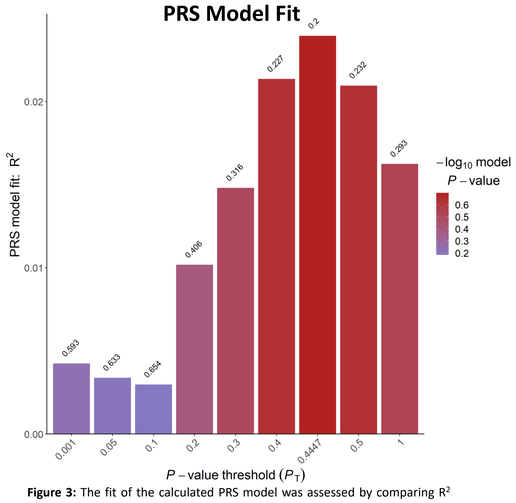

Figure 3: The fit of the calculated PRS model was assessed by comparing R^2 values depending on the P-value threshold used for calculation. P-values of the model at each threshold point are represented by color. This model used only Caucasian samples and alcohol-based questions.

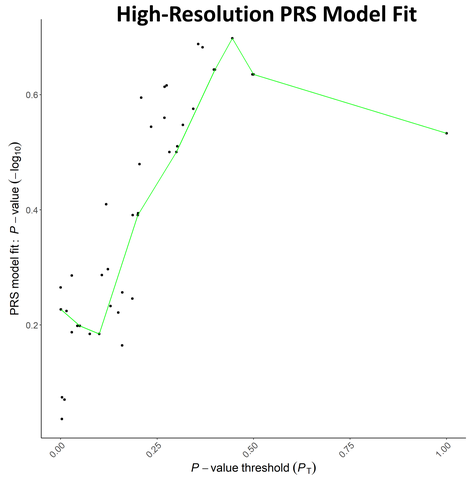

Figure 4: Similarly to Figure 3, the P-value of a given sample is visualized here relative to the P-value threshold used. This model used only Caucasian samples and alcohol-based questions.

Conclusions

The PRS models generated at this phase of the study have not found a statistically significant fit. The score model calculated using only Caucasian samples and alcohol-focused questions appeared to fit the best, but still had low success. Similarly, a slight upward trend can be observed in this Quartile plot (**Figure 2**), but the R^2 and *p* values shown in **Figures 3** and **4** indicate no statistical support. This can likely be attributed to the severe cutting of sample and SNP amounts during QC checks, and differences in analysis methods, population groups, and conditions between this study and the GWAS data analysis. Further studies, including additional sample analyzes and SNP panel design, may minimize these issues. GWAS data from additional population groups could also potentially lead to increased significance of calculated models. Choi et al. (2020) warned in their paper and tutorial accompanying PRSice that biallelic SNPs tend to give extremely low weights and need a fairly large number of variants to be impactful, and that rigorous QC is essential to ensure the data are appropriate for statistical analysis. PRS analysis methods and uses are still being developed, and their value as research continues to advances is yet to be seen.

Genetic Analysis

Buccal swabs were obtained from 525 survey participants. As part of the survey, 69 self-reported as African American, 8 as Asian, 190 as Caucasian, 1 as Pacific Islander, 191 as Hispanic, 3 as American Indian, 7 as other, and 22 did not respond. Genetic sequencing of 158 Single Nucleotide Polymorphisms (SNPs) of interest was performed using a CleanPlex® Custom NGS Panel (Paragon Genomics) on the MiSeq FGx™ instrument (Verogen). Fastq files were converted to Genotype CSV files using the DNA Amplicon tool on the Illumina® BaseSpace Sequence Hub. The data were then compiled and formatted for subsequent analysis using SPSS Statistics® (IBM).

Data Analysis

Survey and genetic data were combined and coded in SPSS. Plink v1.9 was used for QC testing of the genetic data, according to the recommendations provided by Marees et al. (2018). SNPs and individuals were checked for missingness, sex discrepancies, minor allele frequencies, Hardy-Weinberg equilibrium, heterozygosity, relatedness, and population stratification (**Figure 1**), with nonconforming elements removed. PRSice was used to develop PRS models and statistical analysis. Publicly available GWAS data from Sanchez-Roige et al. (2019) were used as a base for the PRS calculation. That study looked at genetic associations with the Alcohol Use Disorders Identification Test (AUDIT) in two large European cohorts (N=121,604 and 20,328). Thus, several PRS calculations were performed for various subsets of the target sample group and survey data, including Caucasian+Alcohol Only, Caucasian+All Substance, All Ancestries+Alcohol Only, and All Ancestries +All Substances.

Acknowledgements

The authors would like to thank the SHSU Departments of Forensic Science and Criminal Justice, especially the efforts of Meghan Royle, Katherine Perez, and Martha Chumchal. Thank you also to the designers and programmers of the various freely accessible programs used in this work, the authors of their comprehensive tutorials. Thank you to Sanchez-Roige et al. for providing their GWAS analysis data publicly, and lastly to all those who have been extremely patient and accommodating during the long completion of this poster.