

Empirical Assessment of Substance Use Disorders Phenotype Strategies for Maximizing Genetic Discovery Power in Biobanks

Amy Moore¹, Amanda E. Gentry², Mohammed F. Hassan², Roseann E. Peterson^{3,1}, Madhurbain Singh², Brien P. Riley², Silviu-Alin Bacanu², Tan Hoang Nguyen², Julie D. White¹, Dana B. Hancock¹, Sandra Sanchez-Roige^{4,5}, Nathan C. Gaddis¹, Olivia Corradin⁵, Brion S. Maher⁶, Abraham A. Palmer^{7,8}, Elissa J. Chesler⁹, Jason A. Bubier⁹, Daniel A. Jacobson¹⁰, Lea K. Davis¹¹, Vanessa Troiani¹², Eric O. Johnson¹, Bradley T. Webb^{1,2}

¹GenOmics and Translational Research Center, RTI International, RTI International, Research Triangle Park, North Carolina, USA; ²Department of Psychiatry, Virginia Institute for Psychiatric and Behavioral Genetics, Virginia Commonwealth University, Richmond, Virginia, USA ; ³Department of Psychiatry and Behavioral Sciences, Institute for Genomics in Health, SUNY Downstate Health Sciences University, Brooklyn, New York, USA; ⁴Department of Psychiatry, University of California San Diego, La Jolla, California, USA; ⁵Whitehead Institute for Biomedical Research, Massachusetts Institute of Technology, Cambridge, MA, USA; ⁶Department of Mental Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA; ⁷Department of Psychiatry, University of California San Diego, La Jolla, CA; ⁸Institute for Genomic Medicine, University of California San Diego, La Jolla, CA; ⁹The Jackson Laboratory, Bar Harbor, ME, USA; ¹⁰Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, TN, USA; ¹¹Icahn School of Medicine at Mount Sinai, New York, NY, USA; ¹²Geisinger Clinic, Geisinger, Danville, Pennsylvania, USA

Molecular genetic investigations of substances use disorders (SUDs) should be large and deeply phenotyped. However, samples meeting both criteria are rare and other phenotyping approaches may yield heritable phenotypes showing moderate genetic correlations with more rigorously assessed phenotypes. While interpreting results across phenotyping strategies may be challenging, increased sample size may offset the power decline often accompanying misclassification. We previously demonstrated that using multiple domains of evidence can a) increase the number of likely cases identified, b) identify individuals at increased risk of SUDs, and c) better classify screened controls for both opioid use disorder (OUD) and alcohol use disorder (AUD) in All of Us.

We empirically evaluated multiple case-control assignment strategies to find the best balance of characteristics including phenotyping effort, sample size, heritability, and expected discovery power. Strategies include using a single domain, such as diagnostic codes or self-reported data to multidomain classifications with strictly screened controls. For OUD, we demonstrated maximal heritability when using strictly defined controls. While a greater ratio of controls to cases can be obtained when using unscreened (7.0-9.9x) versus screened (2.1-2.8x) controls, the expected decreases in detectable allele association sizes are modest (1.087 to 1.079). In contrast to OUD, we did not observe meaningful differences in heritability for AUD when using different control definitions. Within OUD and AUD, the effect of different control definitions on heritabilities were similar across ancestry groups. Further research is needed to optimize cross OUD-AUD and polysubstance analyses in All of Us and other biobanks.