Supplementary Methods

Using Instrumental Variables to Measure Causation over Time in Cross-lagged Panel Models

Madhurbain Singh[1,2,3], Brad Verhulst[4], Philip Vinh[1,2], Yi (Daniel) Zhou[5,2], Luis F. S. Castro-de-Araujo[2], Jouke-Jan Hottenga[3,6], René Pool[3,6], Eco J. C. de Geus[3,6], Jacqueline M. Vink[7], Dorret I. Boomsma[3,6], Hermine H. M. Maes[1,2], Conor V. Dolan[3,6,8], and Michael C. Neale[5,2,1,3,8]

[1] Department of Human and Molecular Genetics, Virginia Commonwealth University, Richmond, Virginia, USA

[2] Virginia Institute for Psychiatric and Behavioral Genetics, Virginia Commonwealth University, Richmond, Virginia, USA

[3] Department of Biological Psychology, Vrije Universiteit, Amsterdam, The Netherlands

[4] Department of Psychiatry and Behavioral Sciences, Texas A&M University, College Station, Texas, USA

[5] Department of Psychiatry, Virginia Commonwealth University, Richmond, Virginia, USA

[6] Amsterdam Public Health Research Institute, Amsterdam, The Netherlands

[7] Behavioural Science Institute, Radboud University, Nijmegen, The Netherlands

[8] Joint last authors

October 04, 2023

Empirical Application of IV-CLPM: Smoking and Alcohol Use

In this empirical example, we analyzed data from the Netherlands Twin Register (NTR; Ligthart et al., 2019) to examine the causal influences between smoking status and alcohol consumption (drinks per week) using both the CLPM and the IV-CLPM models. In the latter model, we used relevant genetic variants as the IV for either trait. In the current analyses, we included genotyped European-ancestry adult individuals with two waves of survey data: adult NTR (ANTR) Survey 8 (collected in 2009) and ANTR Survey 10 (collected in 2012). These two surveys are hereafter referred to as waves 1 and 2, respectively. To avoid the clustering of study participants within families, we selected one individual per family for the current analyses. Thus, we analyzed data from 4,895 individuals, consisting of 1,745 males and 3,150 females (self-reported gender, matched with biological sex inferred from the genotype).

Genotyping

NTR DNA samples included in the current project were genotyped on 3 SNP microarray platforms, namely Affymetrix 6.0 (N= 1984), Affymetrix Axiom (N= 311), and Illumina GSA NTR array (N= 2600). Genotype calling was done following the manufacturer's protocols and white papers. For each platform, DNA samples were QCed using PLINK (Purcell et al., 2007) based on the following:

i. A mismatch between self-reported and biological sex inferred from genotype.
ii. Heterozygosity (F-values between -0.10 and 0.10 considered acceptable).
iii. A mismatch between identity-by-descent (IBD) estimated in Plink and the degree of relatedness expected from the known pedigree.
iv. Sample call rate, with each sample required to have at least 90% of the total SNPs genotyped, plus at least 80% of the SNPs on each chromosome (chr 1-22 and X) successfully genotyped.

SNP QC was based on the following filters (applied separately to each platform): call rate >95%, Hardy-Weinberg Exact (HWE) test p-value > 0.0001, minor allele frequency (MAF) > 0.01, and Mendelian error rate <1%. In addition, based on several plate control samples in Affymetrix 6 (n=4 typed 38-84 times) and Axiom (n=2 typed 33-37 times), SNPs were removed if the genotypes differed more than 1% between these multiple measurements.

As the genotyped data were to be imputed against the European (EUR) super-population of the 1000 Genomes Project Phase-3 (KGP3; Auton et al., 2015), the HRC 1.1 (Haplotype Reference Consortium, 2016; Ega version), and the GoNL (Genome of the Netherlands Consortium, 2014) reference panels, the genotyped SNPs were aligned to the positive strand of *Genome Reference Consortium Human Build 37* (GRCh37) as follows. A combined reference map was first generated based on the SNPs overlapping in the three reference panels, with MAF being less than 0.20 apart across the three reference panels. Then, across the three reference panels, the $(minimum\ MAF + maximum\ MAF)/2$ allele frequency (Fr) was calculated for each SNP. This Fr value was then used to select the SNPs in the NTR if the MAF in the NTR was within +/-0.10 of Fr. Finally, we removed palindromic SNPs with an allele frequency of 0.40-0.60.

After this alignment of the genotypic data to an external reference panel, the three genotyping platforms are also inherently aligned with each other. Therefore, at this point, the data of the three platforms were merged into a single dataset, and based on the overlapping SNPs, IBD was rechecked against the known family structure (now across platforms). Invalid NTR samples (controls), withdrawn consents, and overlapping samples across the three arrays were removed.

After these QC steps, the current analytic sample (N = 4,895) comprised 1,984 individuals with 534,405 SNPs on Axiom, 311 individuals with 537,992 SNPs on Affymetrix 6.0 and 2,600 individuals with 481898 SNPs on Illumina GSA.

Imputation

Before imputation, for each of the three genotype platforms, the SNP name and reference allele were aligned for the three reference panels, and data were converted to VCF format with PLINK. The data were then imputed against the KGP v5, GONL, HRCega, and HRCega+GONL reference panels. Since GPDR restricts us from using imputation servers, the HRC panel of the Ega Sanger website was used. This panel misses the Sardinia, Gecco, and GONL cohorts. We re-added the GONL cohort, selecting only samples present on all chromosomes for both references, renaming SNP names to HRC, and filtering SNPs with minor allele count (MAC) >5 following the procedures described in Haplotype Reference Consortium (2016). After imputation with the Beagle 5.4 software (Browning et al., 2018), the resulting VCF data of the three platforms were

merged into single chromosome sets 1-22 plus X using BCFtools (Li, 2011) for each reference panel. With QCtool version 2.20 (https://www.well.ox.ac.uk/~gav/qctool_v2/), these data were then converted to BGEN format, as well as best guess genotypes using PLINK1.9 (Chang et al., 2015).

Principal Components Analysis

Twenty KGP3 principal components (PCs) for the genotype data were calculated using the SmartPCA tool in the EIGENSTRAT software (Price et al., 2006). For the principal components analysis (PCA), we selected the genotyped SNPs that passed QC present in one of the three platforms from the KGP3 imputed data (as the overlap between platforms is too small to take only genotyped SNPs). These SNPs were then filtered to have MAF >0.05, HWE p >0.001, call rate >0.98, Mendelian error rate < 1%, and imputation info >=90%. These SNPs were subsequently pruned with PLINK (option `–indep 50 5 2`), and SNPs in long-range LD blocks were removed (Price et al., 2008). This left 110,558 SNPs for analysis. From the 1KGP3 reference panel, all samples with the same SNPs were selected and then merged with the NTR data. Subsequently, PCs were calculated in the KGP3 subset and then projected upon the NTR data with the SmartPCA software.

Polygenic Scoring

In this study, we used the results from large-scale European-ancestry GWAS meta-analyses of "smoking initiation" and "drinks per week" (Saunders et al., 2022), excluding the NTR from the GWAS meta-analysis, to derive polygenic scores associated with the smoking status and drinks per week, respectively, in the NTR.

For polygenic scoring, we used the NTR data imputed to the HRC+GONL reference panel. Before scoring, a post-imputation SNP QC selection was employed. This included the following SNP filters: MAF >0.01, HWE p >0.00001, Mendel error rate < 1%, and genotype call rate over 98%. This selection was made on the merged best-guess three-platform data. Furthermore, the imputation info for the three platforms needed to be above 0.10, and the allele frequency between platforms after imputation could not differ more than 2%. This left 7,551,860 SNPs for analysis. Since the GWAS summary statistics are based on KGP3 instead of HRC, we made an NTR reference map to rename SNP IDs back to their respective KGP3 IDs.

The PGSs were calculated using *LDpred* v0.9 (Vilhjálmsson et al., 2015). For estimating the target LD (linkage disequilibrium) structure, we (1) used a selection of unrelated individuals in the NTR sample and (2) selected a set of well-imputed variants in the NTR sample. The parameter `ld_radius` was set by dividing the number of variants in common (from the output of the coordination step) by 12000. Note that for the coordination step, we provided the median sample size as the input value for N. For the LDpred step, we applied the following thresholds for the fraction of variants with non-zero effects (in addition to the default infinitesimal model): `--PS=0.5,0.3,0.2,0.1,0.05,0.01`.

To determine the LDPred threshold that yielded the PGS with the highest predictive power for our outcome variable of interest, we estimated incremental R-squared using a two-step process. We first fitted a null regression model with a standard set of covariates comprising age, sex, SNP microarray platform, and the first ten genetic PCs (without including a PGS). Then, we fitted a full model with a particular PGS as an additional independent variable. The difference in the R-squared estimates of the two models provided the variance in the outcome variable explained by the PGS (controlling for the covariates). For the ordinal variable of smoking status, we fitted ordinal logistic regression models (using the `polr` function in the MASS package (Venables & Ripley, 2002) in R (R Core Team, 2021), and estimated the associated R-squared on the liability scale (Lee et al., 2012). For either outcome variable (smoking status and drinks per week), the PGS with the highest incremental R-squared was retained for further analyses. Accordingly, the PGSs used in the IV-CLPM model were based on a threshold of 0.3 for smoking status and 0.1 for drinks per week. Both PGSs were residualized for the SNP microarray platform and the first ten genetic PCs, and the residuals were then standardized to have a mean of zero and S.D. of one.

.

References

Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., Clark, A. G., Donnelly, P., Eichler, E. E., Flicek, P., Gabriel, S. B., Gibbs, R. A., Green, E. D., Hurles, M. E., Knoppers, B. M., Korbel, J. O., Lander, E. S., Lee, C., . . . Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68-74. https://doi.org/10.1038/nature15393

Browning, B. L., Zhou, Y., & Browning, S. R. (2018). A One-Penny Imputed Genome from Next-Generation Reference Panels. *The American Journal of Human Genetics*, *103*(3), 338-348. https://doi.org/10.1016/j.ajhg.2018.07.015

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience*, *4*(1). https://doi.org/10.1186/s13742-015-0047-8

Genome of the Netherlands Consortium. (2014). Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nature Genetics*, *46*(8), 818-825. https://doi.org/10.1038/ng.3021

Haplotype Reference Consortium. (2016). A reference panel of 64,976 haplotypes for genotype imputation. *Nature Genetics*, *48*(10), 1279-1283. https://doi.org/10.1038/ng.3643

Lee, S. H., Goddard, M. E., Wray, N. R., & Visscher, P. M. (2012). A Better Coefficient of Determination for Genetic Profile Analysis. *Genetic Epidemiology*, *36*(3), 214-224. https://doi.org/10.1002/gepi.21614

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, *27*(21), 2987-2993. https://doi.org/10.1093/bioinformatics/btr509

Ligthart, L., van Beijsterveldt, C. E. M., Kevenaar, S. T., de Zeeuw, E., van Bergen, E., Bruins, S., Pool, R., Helmer, Q., van Dongen, J., Hottenga, J.-J., van't Ent, D., Dolan, C. V., Davies, G. E., Ehli, E. A., Bartels, M., Willemsen, G., de Geus, E. J. C., & Boomsma, D. I. (2019). The Netherlands Twin Register: Longitudinal Research Based on Twin and Twin-Family Designs. *Twin Research and Human Genetics*, *22*(6), 623-636. https://doi.org/10.1017/thg.2019.93

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, *38*(8), 904-909. https://doi.org/10.1038/ng1847

Price, A. L., Weale, M. E., Patterson, N., Myers, S. R., Need, A. C., Shianna, K. V., Ge, D., Rotter, J. I., Torres, E., Taylor, D., Kent, Goldstein, D. B., & Reich, D. (2008). Long-Range LD Can Confound Genome Scans in Admixed Populations. *The American Journal of Human Genetics*, *83*(1), 132-135. https://doi.org/10.1016/j.ajhg.2008.06.005

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., De Bakker, P. I. W., Daly, M. J., & Sham, P. C. (2007). PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics*, *81*(3), 559-575. https://doi.org/10.1086/519795

R Core Team. (2021). *R: A Language and Environment for Statistical Computing*. In R Foundation for Statistical Computing. https://www.R-project.org/

Saunders, G. R. B., Wang, X., Chen, F., Jang, S.-K., Liu, M., Wang, C., Gao, S., Jiang, Y., Khunsriraksakul, C., Otto, J. M., Addison, C., Akiyama, M., Albert, C. M., Aliev, F., Alonso, A., Arnett, D. K., Ashley-Koch, A. E., Ashrani, A. A., Barnes, K. C., . . . Vrieze,

S. (2022). Genetic diversity fuels gene discovery for tobacco and alcohol use. *Nature*, *612*(7941), 720-724. https://doi.org/10.1038/s41586-022-05477-4

Venables, W., & Ripley, B. D. (2002). *Statistics Complements to Modern Applied Statistics with S* (4th ed.). Springer. https://www.stats.ox.ac.uk/pub/MASS4/

Vilhjálmsson, J., Bjarni, Yang, J., Finucane, K., Hilary, Gusev, A., Lindström, S., Ripke, S., Genovese, G., Loh, P.-R., Bhatia, G., Do, R., Hayeck, T., Won, H.-H., Kathiresan, S., Pato, M., Pato, C., Tamimi, R., Stahl, E., Zaitlen, N., Pasaniuc, B., . . . Zheng, W. (2015). Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. *The American Journal of Human Genetics*, *97*(4), 576-592. https://doi.org/10.1016/j.ajhg.2015.09.001