School of Psychological Science
12a Priory Road
Bristol BS8 1TU
United Kingdom

zoe.reed@bristol.ac.uk
www.bristol.ac.uk

ANALYSIS PLAN

What are "smoking initiation SNPs" capturing? Exploring pleiotropy in genetic analyses of smoking-related exposures

Zoe E. Reed, Jasmine N. Khouja, Robyn E. Wootton, Tom G. Richardson, George Davey Smith, Marcus R. Munafò

Background

Recent studies examining different aspects of smoking behaviour suggest that some of the associations may be due to pleiotropic effects (specifically horizontal pleiotropy, where a genetic variant may influence different phenotypes via independent pathways). For example, polygenic risk scores for smoking initiation have been found to be associated with risk taking and externalising behaviours, even in young children (Khouja et al., 2020; Liu et al., 2019; Schellhas et al., 2020), which suggests that the single nucleotide polymorphisms (SNPs) associated with smoking initiation may be capturing phenotypes other than smoking *per se*.

Genome wide association studies (GWAS) are intended to identify SNPs associated with the phenotype being tested. These SNPs can then be used in analyses such as Mendelian Randomisation (MR) (Davey Smith and Ebrahim, 2003; Davey Smith and Hemani, 2014). This assumes that genome-wide significant SNPs are truly associated with the exposure of interest. However, as sample sizes for GWAS have increased (Mills and Rahal, 2019), it is likely that genome-wide significant SNPs are no longer only associated with the trait of interest, but also with other correlated phenotypes, resulting in horizontal pleiotropy.

If SNPs associated with exposures such as smoking initiation are in fact horizontally pleiotropic, this poses a problem for MR analyses that are based on the assumption that SNPs from GWAS are specifically associated with the exposure and that the SNPs are not associated with confounders. Horizontal pleiotropy is one of the mechanisms by which confounding can be reintroduced and if this occurs then the instrument violates the assumptions of MR and is considered invalid (Davey Smith, 2010; Davey Smith et al., 2008), Therefore, it is important to

ascertain whether this is the case for SNPs we use as instruments in MR.

Although several MR sensitivity methods already exist that can help us to infer the likelihood of bias from horizontal pleiotropy, we cannot directly test for pleiotropic effects (without knowledge of the functional biology). Therefore, it is important to interrogate the extent of pleiotropy in genetic instruments for smoking behaviours in other ways, such as those used in the present study. This will help in the interpretation of future MR studies using genetic instruments for smoking behaviours. This is the focus of the current study.

Study Aims

We aim to assess whether the SNPs identified in GWAS for smoking related traits are also associated with correlated confounders, independently of smoking.

Specifically, we will investigate whether the positive and negative control variables we have pre-selected from a PheWAS of smoking initiation (using SNPs from a different sample) can be predicted by genetic risk scores of exposures for smoking related phenotypes of:

- Smoking initiation
- Smoking heaviness
- Lifetime smoking index

We hypothesise that the SNPs in the smoking related GWAS will be associated with confounders as well, which may provide evidence of horizontal pleiotropy.

We will also conduct sensitivity analyses to assess whether the p-value threshold used to create the polygenic risk score impacts our results.

If we find evidence of horizontal pleiotropy in this study, we would follow this up with two further studies.

First, we would assess whether we see evidence of this in a separate cohort – the Million Veteran Program (MVP). Participation in MVP is likely to have different selection biases to participation in UK Biobank, and this may influence the SNPs identified for smoking initiation. Therefore, in turn, this may help us to better understand the reason behind any pleiotropic effects we observe in this study.

Second, we would assess whether we see similar results – where comparable phenotypes are available – in children at an age where they would not yet have smoked (around age 7) in the Avon Longitudinal Study of Parents and Children (ALSPAC). If we also observe similar results here, this would provide further evidence of pleiotropy.

Study Design

We will assess whether the SNPs identified in GWAS for smoking related traits are also picking up other phenotypes (horizontal pleiotropy), using existing data from the UK Biobank and genetically informed analyses. We will use genetic and phenotypic

data from the UK Biobank, a large population-based prospective health research resource of around 500,000 participants, recruited between 2006 and 2010 from across the UK (Sudlow et al., 2015).

Participants

The UK Biobank includes participants aged between 38 and 73 years with data on a range of sociodemographic, lifestyle, physical and mental health measures. Data have been collected via a number of methods, including paper and web-based questionnaires, computer assisted interviews, clinic visits and data linkage. Baseline assessment took place at 22 assessment centres in the UK to enable recruitment from a range of locations, but further data collection is ongoing. Further information can be found on the UK Biobank website (www.ukbiobank.ac.uk). There are 488,377 participants with genotype data available and details pre-imputation quality control, phasing and imputation, as well as in-house quality control filtering have been described elsewhere (Bycroft et al., 2018; Mitchell et al., 2019). We will include participants who have genetic and phenotypic data available for our exposures and outcomes of interest. Participation in UK Biobank is voluntary, and participants are free to withdraw from the cohort at any time without giving a reason. We will exclude any participants who have withdrawn their consent for their data to be used by using the latest withdrawal lists provided for the project data we have access to (UK Biobank project number: 16729). We will also restrict analyses to individuals of "White British" ancestry and remove related individuals.

Measures

Exposures: Smoking initiation, smoking heaviness and lifetime smoking index.

Outcomes: We initially conducted a phenome wide association study (PheWAS) (Denny et al., 2010) for smoking initiation using a polygenic risk score of smoking initiation as the exposure, constructed in UK Biobank. To avoid sample overlap, we used GWAS summary statistics from the GWAS and Sequencing Consortium of Alcohol and Nicotine use (GSCAN) GWAS (Liu et al., 2019) for smoking initiation excluding the UK Biobank sample. We used genome-wide significant SNPs only in our polygenic risk score. The PheWAS was conducted using the PHEnome Scan ANalysis Tool (PHESANT) software package (Millard et al., 2018), which performs phenome scans on data from UK Biobank. We used this to test the association of our polygenic risk score with all of these outcomes (21,409 variables). Of these, 566 variables were associated with the polygenic risk score (at a Bonferroni adjusted p-value threshold of 2.34×10^{-06}). From the top 100 of these (a threshold we decided *a priori*) we selected variables to be positive and negative controls in these analyses (our outcome variables). We did not include those related to the main smoking phenotypes and for similar variables we selected the one that we believed captured the most information. Positive controls were those known (or strongly believed) to be causally related to smoking initiation. Negative controls are those that we consider to be less plausibly causally related to smoking initiation. The effect estimate from the PheWAS are listed next to each variable to indicate the direction of the association.

For positive controls we selected (N=15):

- Body mass index (BMI) (*b*=0.03)
- Body fat percentage (*b*=0.02)
- Wheeze or whistling in the chest in last year (*b*=0.07)
- C-reactive protein (*b*=0.03)
- Date first reported (other chronic obstructive pulmonary disease) * (*b*=0.15)
- Mouth/teeth dental problems: Dentures (*b*=0.07)
- Overall health rating (*b*=0.05, higher value corresponds to poorer health)
- Gamma glutamyltransferase (*b*=0.02)
- White blood cell (leukocyte) count (*b*=0.02)
- Mean sphered cell volume (*b*=0.02)
- Townsend deprivation index at recruitment (*b*=0.03)
- Seen doctor (GP) for nerves, anxiety, tension or depression (*b*=0.05)
- Number of treatments/medications taken (*b*=0.04)
- Father's age at death (*b*=-0.02)
- Amount of alcohol drunk on a typical drinking day (*b*=0.07)

* converted to a binary variable for first occurrence

For negative controls we selected (N=12):

- Lifetime number of sexual partners (*b*=0.08)
- Age at first live birth * (*b*=-0.03)
- Leisure/social activities: Religious group (*b*=-0.07)
- Cereal intake (*b*=-0.04)
- Risk taking (*b*=0.05)
- Time spent watching television (*b*=0.04)
- Liking for cabbage (*b*=0.06)
- Weekly usage of mobile phone in last 3 months (*b*=0.04)
- Ease of skin tanning ** (*b*=-0.03)
- Mother's age at time of questionnaire (*b*=-0.02)
- Pain type(s) experienced in last month (back pain) (*b*=0.04)
- Had an operation on the left-side of the body (*b*=0.04)

* opposite to any adverse effect on fertility
** higher value corresponds to less likely to tan

Confounders: age, sex, first 10 principal components from principal components analysis of the genotype data in UK Biobank.

Statistical Analysis Plan

We will use a 10-fold cross validation approach in an attempt to reduce bias when the samples used for the GWAS and polygenic risk score construction are the same (Burgess et al., 2017). We will use data from the UK Biobank for the smoking related variables of: i) smoking initiation, ii) smoking heaviness, and iii) lifetime smoking index and run 10 GWAS for each smoking phenotype, where each GWAS includes a

different, random sample with 10% of the sample removed.

The polygenic risk scores will then be constructed for this remaining 10% of the sample, to avoid sample overlap. This will result in each participant having a polygenic risk score after all 10 iterations.

We will use the resulting polygenic risk scores and test the association of these with the positive and negative controls from the PheWAS. If we see associations with phenotypes such as risk-taking behaviour or if the three smoking phenotypes show very different associations then this provides evidence of horizontal pleiotropy and may suggest that confounding has been reintroduced.

Sensitivity analyses

We will investigate whether any pleiotropic effects we observe in the main analyses may be due to the inclusion of more SNPs, from larger GWAS, in our instruments for smoking related traits. This may allow us to better understand the profile and extent of pleiotropy for these instruments which will be useful in future MR studies. To do this we will run analyses where we found evidence of an association in our main analyses using different, more stringent p-value thresholds for genome-wide significant SNPs to create the polygenic risk score for the exposure. Specifically, we will run analyses with polygenic risk scores constructed at different p-value thresholds below $5x10^{-08}$ and compare results with those from all genome-wide significant SNPs in the main analyses.

Ethics

UK Biobank received ethics approval from the Research Ethics Committee (REC reference for UK Biobank is 11/NW/0382).

Data Access and Sharing

Phenotypic data from UK Biobank is stored in a project specific folder with access granted only to those on the project on a secure server. Linker ID's with UK Biobank genetic data can be created for each project and linked to genetic data. Any withdrawals of consent are updated by the project lead on a specific project for UK Biobank. All data access will be via a remote server. We will adhere to the relevant data protection legislation, including the EU General Data Protection Regulation (https://www.eugdpr.org/) and UK Data Protection Act 2018. All data provided by UK Biobank is anonymised by a unique identifier.

To access UK Biobank data, researchers must complete an application for a proposed project. Once approved a material transfer agreement will need to be executed before data is released. Further information on data access can be found here (https://www.ukbiobank.ac.uk/using-the-resource/).

The code used for data analysis will be made available upon publication on the data.bris research data repository (https://data.bris.ac.uk/data/).

The results from this study will be published in an appropriate scientific journal (and

made available open access) and/or presented at an appropriate scientific meeting.

Study Personnel

Zoe E. Reed
School of Psychological Science
University of Bristol
12a Priory Rd
Bristol BS8 1TU
Email: zoe.reed@bristol.ac.uk

Jasmine N. Khouja
School of Psychological Science
University of Bristol
12a Priory Road
Bristol, BS8 1TU
Email: jasmine.khouja@bristol.ac.uk

Robyn E. Wootton
Nic Waals Institute
Lovisenberg Diaconal Hospital
Oslo
Email: robyn.wootton@bristol.ac.uk

Tom G. Richardson
Bristol Medical School
University of Bristol
Oakfield Grove,
Bristol, BS8 2BN
Email: tom.g.richardson@bristol.ac.uk

George Davey Smith
Bristol Medical School
University of Bristol
Oakfield Grove,
Bristol, BS8 2BN
Email: kz.davey-smith@bristol.ac.uk

Marcus R. Munafò
School of Psychological Science
University of Bristol
12a Priory Rd
Bristol BS8 1TU
Email: marcus.munafo@bristol.ac.uk

Funding Source

This work will be supported by the UK Medical Research Council Integrative Epidemiology Unit at the University of Bristol (Grant ref: MC_UU_00011/7).

Conflicts of Interest

None.

References

Burgess, S., Small, D.S., Thompson, S.G., 2017. A review of instrumental variable estimators for Mendelian randomization. Stat. Methods Med. Res.

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., Cortes, A., Welsh, S., Young, A., Effingham, M., McVean, G., Leslie, S., Allen, N., Donnelly, P., Marchini, J., 2018. The UK Biobank resource with deep phenotyping and genomic data. Nature 562, 203–209.

Davey Smith, G., 2010. Mendelian randomization for strengthening causal inference in observational studies: Application to gene × environment interactions. Perspect. Psychol. Sci. 5, 527–545.

Davey Smith, G., Ebrahim, S., 2003. "Mendelian randomization": can genetic epidemiology contribute to understanding environmental determinants of disease? Int. J. Epidemiol. 32, 1–22.

Davey Smith, G., Hemani, G., 2014. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. Hum. Mol. Genet. 23, R89--98.

Davey Smith, G., Timpson, N., Ebrahim, S., 2008. Strengthening causal inference in cardiovascular epidemiology through Mendelian randomization. Ann. Med. 40, 524–541.

Denny, J.C., Ritchie, M.D., Basford, M.A., Pulley, J.M., Bastarache, L., Brown-Gentry, K., Wang, D., Masys, D.R., Roden, D.M., Crawford, D.C., 2010. PheWAS: Demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. Bioinformatics 26, 1205–1210.

Khouja, J.N., Wootton, R.E., Taylor, A.E., Smith, G.D., Munafò, M.R., 2020. Association of genetic liability to smoking initiation with e-cigarette use in young adults. medRxiv 2020.06.10.20127464.

Liu, M., Jiang, Y., Wedow, R., Li, Y., Brazel, D.M., Chen, F., Datta, G., Davila-Velderrain, J., McGuire, D., Tian, C., Zhan, X., Agee, M., Alipanahi, B., Auton, A., Bell, R.K., Bryc, K., Elson, S.L., Fontanillas, P., Furlotte, N.A., Hinds, D.A., Hromatka, B.S., Huber, K.E., Kleinman, A., Litterman, N.K., McIntyre, M.H., Mountain, J.L., Northover, C.A.M., Sathirapongsasuti, J.F., Sazonova, O. V., Shelton, J.F., Shringarpure, S., Tung, J.Y., Vacic, V., Wilson, C.H., Pitts, S.J., Mitchell, A., Skogholt, A.H., Winsvold, B.S., Sivertsen, B., Stordal, E., Morken, G., Kallestad, H., Heuch, I., Zwart, J.A., Fjukstad, K.K., Pedersen, L.M., Gabrielsen, M.E., Johnsen, M.B., Skrove, M., Indredavik, M.S., Drange, O.K., Bjerkeset, O., Børte, S., Stensland, S.Ø., Choquet, H., Docherty, A.R., Faul, J.D., Foerster, J.R., Fritsche, L.G., Gordon, S.D., Haessler, J., Hottenga, J.J., Huang, H., Jang, S.K., Jansen, P.R., Ling, Y., Mägi, R., Matoba, N., McMahon, G., Mulas, A., Orrù, V., Palviainen, T., Pandit, A., Reginsson, G.W., Smith, J.A., Taylor, A.E., Turman, C., Willemsen, G., Young, H., Young, K.A., Zajac,

G.J.M., Zhao, W., Zhou, W., Bjornsdottir, G., Boardman, J.D., Boehnke, M., Boomsma, D.I., Chen, C., Cucca, F., Davies, G.E., Eaton, C.B., Ehringer, M.A., Esko, T., Fiorillo, E., Gillespie, N.A., Gudbjartsson, D.F., Haller, T., Harris, K.M., Heath, A.C., Hewitt, J.K., Hickie, I.B., Hokanson, J.E., Hopfer, C.J., Hunter, D.J., Iacono, W.G., Johnson, E.O., Kamatani, Y., Kardia, S.L.R., Keller, M.C., Kellis, M., Kooperberg, C., Kraft, P., Krauter, K.S., Laakso, M., Lind, P.A., Loukola, A., Lutz, S.M., Madden, P.A.F., Martin, N.G., McGue, M., McQueen, M.B., Medland, S.E., Metspalu, A., Mohlke, K.L., Nielsen, J.B., Okada, Y., Peters, U., Polderman, T.J.C., Posthuma, D., Reiner, A.P., Rice, J.P., Rimm, E., Rose, R.J., Runarsdottir, V., Stallings, M.C., Stančáková, A., Stefansson, H., Thai, K.K., Tindle, H.A., Tyrfingsson, T., Wall, T.L., Weir, D.R., Weisner, C., Whitfield, J.B., Yin, J., Zuccolo, L., Bierut, L.J., Hveem, K., Lee, J.J., Munafò, M.R., Saccone, N.L., Willer, C.J., Cornelis, M.C., David, S.P., Jorgenson, E., Kaprio, J., Stitzel, J.A., Stefansson, K., Thorgeirsson, T.E., Abecasis, G., Liu, D.J., Vrieze, S., 2019. Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use. Nat. Genet.

Millard, L.A., Davies, N.M., Gaunt, T.R., Smith, G.D., Tilling, K., 2018. Software application profile: PHESANT: A tool for performing automated phenome scans in UK Biobank. Int. J. Epidemiol. 47, 29–35.

Mills, M.C., Rahal, C., 2019. A scientometric review of genome-wide association studies. Commun. Biol.

Mitchell, R., Hemani, G., Dudding, T., Corbin, L., Harrison, S., Paternoster, L., 2019. UK Biobank Genetic Data: MRC-IEU Quality Control, Version 2.

Schellhas, L., Haan, E., Easey, K., Wootton, R.E., Sallis, H., Sharp, G.C., Munafo, M.R., Zuccolo, L., 2020. Maternal and child genetic liability for smoking and caffeine consumption and child mental health: An intergenerational polygenic risk score analysis in the ALSPAC cohort. medRxiv 2020.09.07.20189837.

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., Liu, B., Matthews, P., Ong, G., Pell, J., Silman, A., Young, A., Sprosen, T., Peakman, T., Collins, R., 2015. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. PLoS Med. 12, e1001779.