Title: Gene-environment interactions in Multiple Sclerosis: a UK Biobank study

Authors: Benjamin Meir Jacobs (BM BCh)[1,2], Alastair Noyce (PhD) [1,2], Jonathan Bestwick (PhD) [1], Daniel Belete (MBChB) [1,2], Gavin Giovannoni (PhD)[1,2], Ruth Dobson (PhD)[1,2]

Affiliations: 1: Preventive Neurology Unit, Wolfson Institute of Preventive Medicine, Barts and Queen Mary University of London. 2: Royal London Hospital, Barts Health NHS Trust.

Corresponding author: Ruth Dobson, Preventive Neurology Unit, Wolfson Institute of Preventive Medicine, Barts and Queen Mary University of London, ruth.dobson@qmul.ac.uk.

Acknowledgements and data availability

We would like to thank the relevant consortia for making their data available. MS GWAS data were taken from the MS Chip discovery summary statistics. IMSGC summary statistics are not publicly available and were obtained via request on the website: https://nettskjema.no/answer/imsgc-data-access.html. We would like to thank the Queen Mary University High Performance Computing team for their help with computing resources. We would like to thank the participants and researchers involved in UK Biobank, who have created an exceptional resource. UK Biobank data are available on request through their website. Code used in this paper is available on Github (@benjacobs123456).

Abstract

Importance:

Multiple Sclerosis (MS) is a common neuro-inflammatory disorder caused by a combination of environmental exposures and genetic risk factors. Interaction between environmental and genetic factors may impact on MS risk.

Objective:

To determine whether genetic risk modifies the effect of environmental MS risk factors.

Design and setting:

Retrospective case-control study using data from a longitudinal cohort (UK Biobank).

Participants:

People with MS (pwMS; 72.7% female, mean age=55.2, SD=7.64, median age at diagnosis=41.06) were identified using ICD10-coded MS or self-report. The remainder of the cohort was used as controls. For interaction, only people with white British ancestry were included.

Exposure(s):

Confounders: age, sex, Townsend deprivation index at recruitment, self-reported ethnicity, birth latitude. Exposures: age at puberty, age at first sexual intercourse, birth weight, breastfeeding, exposure to maternal smoking, month of birth, smoking status, body size aged 10, and self-reported Infectious Mononucleosis. Genetic exposures were HLA-DRB1*15, HLA-A*02, and an autosomal non-HLA genetic risk score.

Main Outcome(s) and Measure(s):

Associations with MS risk were quantified using odds ratios from multivariable logistic regression. Interaction between environmental and genetic risk factors was quantified using the Attributable Proportion due to interaction (AP). Departure from additivity refers to the

risk of an outcome which exceeds the risk expected from adding individual excess risks (risk differences) together. Model fits were quantified using Nagelkerke's pseudo-R^2 metric.

Results:

Phenotype data were available for 2151 pwMS and 486,125 controls. Exposures associated with MS risk were childhood obesity (OR=1.39, 95%CI 1.22-1.58), smoking (OR=1.19, 95%CI 1.07-1.33), earlier menarche 0.95, 95%CI 0.92-0.98), HLA-DRB1*15 ($OR_{Homozygote}$ 5.05, 95%CI 4.22-6.05) and lack of the HLA-A*02allele ($OR_{Homozygote}$=0.57, 95%CI 0.46-0.70). The autosomal polygenic risk score (PRS) was associated with MS disease status ($OR_{\mathbf{Top\text{-}vs\text{-}bottom\text{-}decile}}$=3.96, 95%CI 3.11-5.04). There was evidence of positive (synergistic) interaction between elevated childhood body size and the PRS (AP 0.11, 95% CI 0.008 to 0.202, p = 0.036), and weaker evidence suggesting a possible interaction between smoking status prior to age 20 and the PRS (AP 0.098, 95% CI -0.013 to 0.194, p = 0.082).

Conclusions and Relevance:

This study provides novel evidence for an interaction between childhood obesity and a high burden of autosomal genetic risk. These findings have significant implications for our understanding of MS biology, and inform targeted planning of prevention strategies.

Introduction

Susceptibility to Multiple Sclerosis (MS) is multifactorial: both a large heritable component[1] and a number of environmental associations[2,3] have been identified through a combination of genetic and epidemiological studies. Established environmental exposures associated with MS include smoking, obesity during adolescence, vitamin D deficiency, increasing latitude, infectious mononucleosis (IM) and Epstein-Barr Virus seropositivity. Additional protective associations of CMV seropositivity and breastfeeding, and harmful associations with HHV6 seropositivity, night shift work, organic solvent exposure, low dietary fatty acids, head injury, earlier puberty, and maternal smoking have been identified[2,3].

The genetic architecture of MS susceptibility has been delineated through the efforts of the International Multiple Sclerosis Genetic Consortium (IMSGC). Meta-analysis of genome-wide association studies in over 47,000 cases and 68,000 controls revealed 233 independent loci, accounting for ~48% of the estimated heritability of MS[1]. Of the total heritability explained by common genetic variation, the Major Histocompatibility Complex (MHC) locus accounts for ~20%, and non-MHC for ~20%[1]. Attempts to model MS risk using polygenic risk scores have had some success[4–6], supporting the view that MS susceptibility is influenced by both MHC and genome-wide variation, that non-genetic factors play a substantial role, and that a substantial proportion of population genetic risk is not explained by common variants[1].

Evidence from Scandinavian and North American cohorts suggests that environmental influences on MS risk can be modified by HLA genotype. The deleterious effects of childhood obesity, smoking, infectious Mononucleosis, and solvent exposure on MS risk are reported to be potentiated among carriers of the HLA DRB1*15 allele and people who do not

carry the protective HLA A*02 genotype[7–9,11]. It is not known whether such gene-environment interactions extend beyond the HLA locus in the pathogenesis of MS.

This work sets out to exploit the availability of deep phenotyping and genotype data in UK Biobank to validate and extend previous case-control studies examining MS susceptibility. In particular, we focus on exposures during early life and adolescence, as these are well-characterised in the UK Biobank population, and are less likely to be confounded by prodromal disease, which may influence behaviour and exposure years before the disease is diagnosed.[12]

Methods

Data sources

UK Biobank is a longitudinal cohort study described in detail elsewhere[13]. In brief, participants between the ages of 40 and 69 were recruited between 2006 and 2010 from across the UK. Participants underwent genotyping, donated body fluid samples, and answered a range of questions about lifestyle, environmental and demographic factors. Health records were linked to participants using Hospital Episode Statistics (HES). Phenotype data are composed of survey data, linked healthcare records, anthropometric measurements, and a variety of other biochemical and imaging data.

Identification of cases and controls

We determined MS status (case vs control) using the following approach: individuals were defined as cases if they had at least one ICD-coded diagnosis of Multiple Sclerosis (ICD10 G35; ICD9 3409) or if they self-reported a diagnosis of MS. ICD codes in UK Biobank are

extracted from HES and refer to diagnoses recorded during a hospital admission/encounter. We included all participants with any MS code in either a main or secondary diagnosis field. Age at diagnosis was determined using the approximate age of diagnosis from self-report. Controls were unmatched UK Biobank participants without a coded diagnosis of MS.

Genotype data

Individuals were genotyped using the Axiom or Bileve arrays. Genotyping and quality control protocols are described in detail elsewhere[14]. Imputed HLA alleles were provided by UK Biobank. HLA alleles were imputed to four-digit resolution using the HLA*IMP:02 software with a multi-population reference. We extracted each participant's allelic dosage for the MS risk allele HLA-DRB1*15:01 and the protective allele HLA-A*02:01 by thresholding posterior allele probabilities at 0.7 as suggested by UK Biobank. These two HLA alleles were used as they have the largest effect sizes across multiple studies[2]**.** The imputation procedures and quality control are described in detail elsewhere[14]. Genetic principal components and kinship coefficients were supplied by UK Biobank.

Construction of a polygenic risk score (PRS)

A variety of PRS (twenty-eight in total) were created using the clumping-and-thresholding approach:

1. We extracted variant associations with MS from the discovery stage meta-analysis summary statistics obtained from the IMSGC[1].
2. We excluded variants within the extended MHC (chr6:25,000,000 to chr6:35,000,000 on hg19), those with strand-ambiguous alleles (A/T and C/G SNPs), and variants without an rsid.

3. We excluded variants with association statistic p values above an arbitrary p value threshold (0.01, 0.1, 0.2, 0.4, 0.6, 0.8, and 1).
4. We clumped using several r^2 thresholds (0.2, 0.4, 0.6, 0.8) and a clumping distance of 250kBP, with the 1000 genomes EUR samples as a reference genome.

Reference genome data were obtained from the 503 participants of European ancestry in the 1000 genomes project[15]. Only autosomal, biallelic variants which passed quality control in both the reference and target (UK Biobank) datasets were included. We excluded all duplicate rsIDs, duplicate positions, variants deviating from Hardy-Weinberg Equilibrium (p <1e-06), rare variants with minor allele frequencies <0.01, variants with genotype missingness <10%, and variants with low imputation quality (R^2< 0.3). For genetic analysis, individuals with >10% missing genotypes were excluded, and only individuals with self-reported 'British' ethnicity and genetic 'Caucasian' ancestry as defined by genetic principal components were included. We excluded one of each pair of related individuals (Kinship coefficient > 0.0844). A total of 486125 controls and 2151 cases were included in the case-control study. After exclusion of related individuals, and restricting to only individuals with both self-reported and genetic white British ancestry, 375986 controls and 1740 cases remained.

Beta coefficients from the IMSGC discovery GWAS were calculated from odds ratios[1]. Standard errors were estimated from odds ratios and p values. Effect allele dosage at each locus was multiplied by the beta coefficient to generate the risk score for that locus. Scores were standardised to have mean 0 and unit variance for each SNP. For missing genotypes, the score at that locus was defined as the mean of all scores at that locus. Risk scores were

totalled across the genome to calculate an individual's score. Analysis was performed in PLINK2 using the '--score' flag[16,17].

In order to examine gene-environment interactions on the genome-wide scale, twenty-eight polygenic risk scores (PRS) for MS were created using the pruning-and-thresholding approach excluding the HLA region (see above). PRS scores were normalised using inverse-rank normal transformation. Depending on the underlying genetic architecture of the trait, the pruning and thresholding parameters which give optimal PRS performance vary[18]. We selected the best-fitting PRS by fitting a logistic regression model with MS status as the outcome, and including age, sex, birth latitude, current deprivation, and the first four genetic principal components as covariates. We evaluated model fit using Nagelkerke's pseudo- R^2 metric, comparing the full model to a null model consisting of age, sex, birth latitude, current deprivation, and the first four genetic principal components as covariates. The PRS with the highest Nagelkerke's pseudo- R^2 metric was used for further analyses[17].

Definition of exposures

All reliably coded exposures pertaining to early life, childhood and adolescence in UK Biobank were used. These included variables previously shown to be associated with MS risk: month of birth, age at menarche, breastfeeding, comparative body size at age 10 (CBS_{10}), smoking before the age of 20, and infectious mononucleosis. Age, ethnicity, sex, birth latitude, and Townsend deprivation index at recruitment were included as additional covariates in all models to control for confounding[19]. Vitamin D status was not included, as only current vitamin D levels are available, which are liable to reverse causation and/or confounding. Where multiple data points were available for a participant, the first recorded

reading was used. Details on exposures are given in supplementary table 1. This yielded a total of twelve exposures selected for analysis.

The number of HES-coded infectious mononucleosis (IM) cases within UK Biobank was very small (79 ICD-10 coded cases). IM prevalence in Scandinavian cohorts is around 8-11%; giving an estimated 50,000 cases in Biobank[20]. The HES code-derived estimate is likely a significant underestimate due to the small proportion of IM cases presenting to secondary care. We therefore used self-reported IM status.

Statistical methods

Multivariable models were built for each risk factor including age, sex, ethnicity, current deprivation status, and birth latitude as confounders, using the entire UK Biobank cohort as controls. A total of twelve models were built using the selection criteria described above. Secondly, a multivariable logistic regression model comprising all environmental factors with robust associations to MS risk was built, including the above confounders. Model likelihood ratio was used to assess the improvement of model fit after correcting for multiple comparisons ($p<0.05/12$ for 12 models).

For interaction analyses, the CBS_{10} variable was dichotomised such that participants were classified as “not overweight” if they answered “thinner” or “average”, and “overweight” if they answered “plumper”. Smoking status was characterised as “ever” or “never” smoking. We treated HLA alleles (DRB1*15 and A*02) as additive traits, with each individual coded as having 0, 1, or 2 alleles. Age at menarche was treated as a continuous variable. All

analyses regarding menarche were restricted to women and models do not include a sex term.

Interaction was assessed on both the additive and multiplicative scales. Interaction on the additive scale was assessed by calculating the Attributable Proportion due to interaction (AP). Additive interaction analyses were based on multivariable logistic regression models incorporating age, sex, and the first four genetic principal components as confounders[21].

For a logistic regression model of the form:

$$log(\frac{p}{1-p}) = \beta_0 + \beta_{RF1}x + \beta_{RF2}y + \beta_{RF1*RF2}x \times y$$

In which *log(p/1-p)* is the log odds of MS, *x* and *y* are the values of exposure variables (e.g. childhood body size, smoking, polygenic risk score), and xy is the interaction term, then the Relative Excess Risk due to Interaction (RERI) can be calculated as:

$$RERI = exp(\beta_{RF1} + \beta_{RF2} + \beta_{RF1*RF2}) - exp(\beta_{RF1}) - exp(\beta_{RF2}) + 1$$

The AP can be conceived of as the proportion of the disease in the doubly-exposed group attributable to the interaction between the risk factors, i.e:

$$AP = \frac{RERI}{exp(\beta_{RF1} + \beta_{RF2} + \beta_{RF1*RF2})}$$

This model can be expanded to include confounding covariates, in which case the beta coefficients are adjusted for confounders[21]. We restricted this analysis to participants with genetically European ancestry determined by both self-report (“Caucasian”) and genetic ethnic grouping.

For interaction analyses using the PRS, covariates were age, sex, birth latitude, Townsend deprivation index, and the first four genetic principal components. For the PRS-menarche interaction analysis, sex was not included as a covariate as the analysis was restricted to females. The PRS was transformed using the inverse-normal transformation and treated as a continuous variable for these analyses. Confidence intervals for the AP were estimated using bootstrap resampling of the entire dataset with replacement for 10000 iterations[21] with 95% confidence intervals derived from the 2.5th and 97.5th centile values. Two-sided p values for the AP due were calculated from the exact method with a correction for finite sampling, i.e. for AP > 0:

$$p = 2 \times \frac{1 + number\ of\ iterations < 0}{10001}$$

Interaction on the multiplicative scale was assessed using a logistic regression model incorporating an interaction term and quantified using the likelihood ratio.

Ethical approval

This work was performed using data from UK Biobank (REC approval 11/NW/0382). All participants gave informed consent on Biobank registration and are free to withdraw from the study at any point, at which point their data are censored and cannot be included in further analyses.

Computing

This research was supported by the High-Performance Cluster computing network hosted by Queen Mary, University of London[22].

Statistical analyses were performed in R version 3.6.1 using RStudio version 1.2.1335. Extraction of European individuals from the 1000 genomes reference genome was conducted using vcftools. Construction of the polygenic risk score, application of the polygenic risk score to individuals, and quality control were performed in PLINK 1.9 and PLINK2.

Results

Population demographics

Phenotype data were available for 488,276 UK Biobank participants comprising 2151 people with MS (pwMS) and 486,125 unmatched controls. Among pwMS, the median age at diagnosis was 41.06 (IQR 14.28). Demographic characteristics of people with pwMS and unmatched controls (entire UK Biobank cohort) are shown in table 1.

Exposures associated with MS in UK Biobank

After adjustment for age, sex, ethnicity, birth latitude and current deprivation, factors associated with increased risk of MS were higher CBS_{10}, smoking, earlier menarche, carriage of the HLA DRB1*15:01 risk allele, and lack of the protective HLA A*02:01 allele (table 2, Fig. 1). The estimate for IM was imprecise (OR 1.70, 95% CI 0.88 to 3.23), likely due to the small number of IM cases in this analysis (Fig. 1). We did not examine interactions between genotype and IM because IM was not clearly associated with MS risk in this analysis.

Gene-environment interactions

To analyse gene-environment interactions we used data from unrelated individuals of European descent, yielding 375986 controls and 1740 cases. We selected the best-fitting PRS

using Nagelkerke's pseudo- R^2metric (methods) (Fig. 2). Participants in the highest score decile of the best performing score were more likely to have MS than those in the lowest decile (OR 3.96, 95% CI 3.11-5.04). The PRS added a modest amount of additional explanatory power to models containing HLA, environmental risk factors (for this analysis, only smoking and CBS_{10} to avoid excluding males), and a combination of the two (Fig. 2), suggesting that the PRS captures genetic risk which is independent of the effects of these other predictors.

There was evidence of positive (synergistic) interaction between elevated childhood body size and the PRS (AP 0.11, 95% CI 0.008 to 0.202, p = 0.036), although this did not survive multiple comparison testing (threshold p = 0.05/9). We found weaker evidence suggesting a possible interaction between smoking status prior to age 20 and the PRS (AP 0.098, 95% CI -0.013 to 0.194, p = 0.082). There was no evidence of additive interaction between age at menarche and the PRS (table 3, Fig. 3), nor was there strong evidence of multiplicative interaction between the PRS and any of these three exposures (data not shown). To illustrate the practical importance of these putative interactions, we performed stratified logistic regression modelling the effect of childhood body size and smoking for individuals in the highest and lowest PRS decile groups. The effects of childhood obesity ($OR_{MS|Overweight\ \&\ high\ PRS}$ 1.40, 95% CI 1.08 - 1.83; $OR_{MS|Overweight\ \&\ low\ PRS}$ OR 1.03, 95% CI 0.56 - 1.83) and smoking status prior to age 20 ($OR_{MS|Smoker\ \&\ high\ PRS}$ 1.48, 95% CI 1.14 to 1.92; $OR_{MS|Smoker\ \&\ low\ PRS}$ 1.11, 95% CI 0.64 to 1.92) on MS risk were more pronounced for the highest PRS decile than the lowest PRS decile (Fig. 3). We were unable to demonstrate strong evidence for interaction between HLA genotype and any of the environmental exposures tested, however the estimates were highly imprecise (Fig. 3, table 3). There was no evidence of statistically significant interaction on the multiplicative scale for any traits (data not shown).

Discussion

In this study we used data from UK Biobank to further our understanding of how gene-environment interactions contribute to MS risk. We demonstrate suggestive evidence of a novel interaction on the additive scale between autosomal genetic risk for MS and childhood body size. In addition, we replicate the associations between smoking, childhood body size, early menarche, and MS risk. We found no clear evidence of interaction on the additive scale between these risk factors and HLA genotype.

Polygenic risk scores (PRS) capture the genetic risk conferred by genome-wide variation. The autosomal PRS in this study - which excluded variation within the MHC - captured a small proportion of overall MS liability, but was robustly associated with MS. Previous efforts using PRS from the IMSGC explained up to ~3% of liability[5]. The best-performing PRS in this study explained ~1% of MS liability. This discrepancy could be explained by several factors, including the relatively low number of cases in UK Biobank, the possibility of missed cases, the possibility of self-report being less accurate than clinical diagnosis, differences in population structure, restriction according to self-declared ethnicity and ethnicity as determined by genetic principal components analysis, and some SNPs not being available/failing QC checks in Biobank. Nevertheless, despite the low overall liability captured, the validity of the PRS is underscored by the monotonic relationship between PRS and OR of MS, the robust model fit when using the PRS to model MS risk, and the disease-specificity of the PRS.

We provide suggestive evidence that the impact on MS risk of elevated childhood body size is intensified among individuals with a high genomic risk of MS. Although these results should be interpreted cautiously in light of the lack of prospective data, they suggest that - on a population level - preventing childhood obesity may prevent greater numbers of MS cases

among high-risk individuals. To our knowledge, this is the first evidence of interaction between an environmental risk factor for MS and genome-wide, non-HLA risk[3]. It should be noted that this PRS, as is common, would not perform well for individual risk prediction given the substantial overlap in score distributions between control participants and pwMS[18]. Our results also suggest that there may be a similar direction of effect for smoking prior to aged 20, however the lack of precision in this estimate warrants caution in interpreting this result. It should be noted that the method we use for estimating p values of the attributable proportion due to interaction (taking the absolute number of replicates above or below 0) is conservative compared to asymptotic tests, which assume that the test statistic follows a distribution (e.g. the z distribution).

Our failure to replicate the observed interactions between HLA genotypes, smoking, and childhood body size[789] could be explained by methodological differences between our study and published literature. Our cohort is likely to differ in key respects from the Kaiser Permanente and EIMS cohorts in that UK Biobank is a participants are predominantly Caucasian, from relatively affluent parts of the UK, are self-selecting and are middle-ages (recruitment from 40-69). We control for different covariates in our interaction analyses (using principal components to account for ancestry) and we used imputed HLA alleles to four-digit resolution. UK Biobank survey data is also prone to recall bias as it is retrospective. Prospective data – as has been used to demonstrate the HLA-environmental interactions - are likely to provide more reliable estimates of gene-environment interactions, and thus we would interpret the lack of HLA-environment interactions in our study with caution, as an absence of evidence rather than evidence of absence.

The limited overall liability explained by the PRS, the relatively small absolute number of people with MS, and the imperfect nature of self-reported phenotypes all limit the interpretation of these results. The inherent biases that come with retrospective survey data, using prevalent rather than incident cases, and the lack of some known important exposures in the dataset (childhood vitamin D levels were not available) are further limitations. MS prevalence in UK Biobank approaches the expected UK prevalence, suggesting that the majority of individuals with MS are correctly identified.

This study thus provides novel suggestive evidence that childhood body size – and possibly smoking - interacts with non-HLA MS genetic risk. It additionally replicates the associations between childhood BMI, smoking, earlier menarche, and MS. Demonstrating benefit for preventive measures in rare, complex diseases like MS is a challenge due to the low population incidence and the small effects of individual interventions. Power can be enhanced by enriching for high-risk individuals, and by selecting individuals who are likely to experience greater benefit from the intervention. As the effects of childhood body size and smoking on MS risk appear greater among individuals with a high genome-wide genetic risk, trials attempting to demonstrate benefit of smoking prevention/cessation may benefit from risk-stratifying individuals using this approach.

References

1. International Multiple Sclerosis Genetics Consortium*†. Multiple sclerosis genomic map implicates peripheral immune cells and microglia in susceptibility. *Science* **365**, eaav7188 (2019).
2. Olsson, T., Barcellos, L. F. & Alfredsson, L. Interactions between genetic, lifestyle and environmental risk factors for multiple sclerosis. *Nat. Rev. Neurol.* **13**, 25–36 (2017).
3. Alfredsson, L. & Olsson, T. Lifestyle and Environmental Factors in Multiple Sclerosis. *Cold Spring Harb. Perspect. Med.* **9**, (2019).
4. Disanto, G. *et al.* The refinement of genetic predictors of multiple sclerosis. *PLoS One* **9**, e96578 (2014).
5. The International Multiple Sclerosis Genetics Consortium (IMSGC). Evidence for Polygenic Susceptibility to Multiple Sclerosis—The Shape of Things to Come. *Am. J. Hum. Genet.* **86**, 621–625 (2010).
6. Dobson, R. *et al.* A Risk Score for Predicting Multiple Sclerosis. *PLoS One* **11**, e0164992 (2016).
7. Hedström, A. K. *et al.* Interaction between passive smoking and two HLA genes with regard to multiple sclerosis risk. *Int. J. Epidemiol.* **43**, 1791–1798 (2014).
8. Hedström, A. K. *et al.* Smoking and two human leukocyte antigen genes interact to increase the risk for multiple sclerosis. *Brain* **134**, 653–664 (2011).
9. Hedström, A. K. *et al.* Interaction between adolescent obesity and HLA risk genes in the etiology of multiple sclerosis. *Neurology* **82**, 865–872 (2014).
10. Hedström, A. K. *et al.* Organic solvents and MS susceptibility: Interaction with MS risk HLA genes. *Neurology* **91**, e455–e462 (2018).
11. Nielsen, T. R. *et al.* Effects of infectious mononucleosis and HLA-DRB1*15 in multiple sclerosis. *Mult. Scler.* **15**, 431–436 (2009).
12. Bjornevik, K. *et al.* Serum Neurofilament Light Chain Levels in Patients With Presymptomatic Multiple Sclerosis. *JAMA Neurol.* (2019). doi:10.1001/jamaneurol.2019.3238
13. Sudlow, C. *et al.* 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 (2015).
14. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
15. Consortium, T. 1000 G. P. & The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
16. Khera, A. V. *et al.* Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat. Genet.* **50**, 1219–1224 (2018).
17. Choi, S. W., Mak, T. S. H. & O'Reilly, P. F. A guide to performing Polygenic Risk Score analyses. *bioRxiv* 416545 (2018). doi:10.1101/416545
18. Janssens, A. C. J. W. & Joyner, M. J. Polygenic Risk Scores That Predict Common Diseases Using Millions of Single Nucleotide Polymorphisms: Is More, Better? *Clin. Chem.* **65**, 609–611 (2019).

19. Murray, S., Bashir, K., Penrice, G. & Womersley, S. J. Epidemiology of multiple sclerosis in Glasgow. *Scott. Med. J.* **49**, 100–104 (2004).
20. Hedström, A. K., Lima Bomfim, I., Hillert, J., Olsson, T. & Alfredsson, L. Obesity interacts with infectious mononucleosis in risk of multiple sclerosis. *Eur. J. Neurol.* **22**, 578–e38 (2015).
21. Knol, M. J., van der Tweel, I., Grobbee, D. E., Numans, M. E. & Geerlings, M. I. Estimating interaction on an additive scale between continuous determinants in a logistic regression model. *Int. J. Epidemiol.* **36**, 1111–1118 (2007).
22. King, T., Butcher, S. & Zalewski, L. *Apocrita - High Performance Computing Cluster for Queen Mary University of London*. (2017). doi:10.5281/zenodo.438045

Figure and table legends

Figures

Figure 1: odds ratios and 95% confidence intervals for the association of each exposure with MS. ORs and Cis are from the output of a multivariable logistic regression with the following covariates: age, sex, ethnicity, birth latitude, current deprivation status, and the exposure in question. For menarche (females only) and voice-breaking (males-only), sex was not included as a covariate.

Figure 2: A: Nagelkerke's pseudo- R^2 metric for each of the individual PRS used. The R^2 was calculated by comparing the model fit (age, sex, birth latitude, Townsend deprivation index, the first 4 genetic PCs, and PRS) vs the null model (age, sex, birth latitude, Townsend deprivation index, and the first 4 genetic PCs). A variety of p value thresholds and clumping parameters were used to create different PRS. Note that the clumping R^2 refers to the linkage disequilibrium threshold within which variants were 'clumped', and is a different quantity from the Nagelkerke pseudo- R^2. B: odds ratios and 95% confidence intervals for MS for individuals in each PRS decile (reference: lowest decile). ORs were calculated from logistic regression models with the following covariates: age, sex, first 4 genetic PCs, and PRS. C: histogram showing PRS distributions among MS cases and controls. D: Nagelkerke pseudo- R^2 metric for models of MS risk. Models were as follows: PRS: MS risk ~ age + sex + first 4 genetic PCs + PRS. HLA: MS risk ~ age + sex + first 4 genetic PCs + HLA genotypes. ENV:

MS risk ~ age + sex + first 4 genetic PCs + childhood BMI + smoking. HLA + PRS: MS risk ~ age + sex + first 4 genetic PCs + HLA genotypes + PRS. HLA+ENV: MS risk ~ age + sex + first 4 genetic PCs + HLA genotypes + childhood BMI + smoking. HLA+PRS+ENV: MS risk ~ age + sex + first 4 genetic PCs + HLA genotypes + PRS + childhood BMI + smoking.

Figure 3: A: Forest plot demonstrating Attributable Proportion due to interaction (AP) and 95% CIs for interactions between environmental exposures and genetic risk factors for MS. If there is no interaction, the AP is 0. AP > 1 indicates positive interaction (combined effects exceed the sum of the individual effects), and vice-versa. APs depicted are derived from logistic regression adjusted for age, sex, and the first 4 genetic principal components. CIs are derived from taking the 2.5^{th} and 97.5^{th} percentiles of 10000 bootstrap replicates. B: forest plot demonstrating odds ratios and 95% CIs for MS given childhood body size (overweight vs not overweight) and smoking status at aged 20 (smoker vs non-smoker). ORs are from the output of logistic regression model of the form MS risk ~ Age + Sex + first 4 genetic PCs. Models were built separately for individuals with the highest 10% of genetic risk scores and the lowest 10% of genetic risk scores ('top' and 'bottom' decile respectively).

Tables

Table 1: demographic characteristics of included participants. Continuous variables are presented as mean(SD), categorical variables are presented as n(%). Missing data are not tabulated. Proportions are calculated as a proportion of individuals with non-missing data for each variable.

Table 2: odds ratios for Multiple Sclerosis for each exposure studied. The first three columns depict the multivariable odds ratios from the output of regression models incorporating age, sex, ethnicity, birth latitude and current deprivation as covariates. Predictors which conferred good model fit (likelihood ratio p value < multiple testing threshold for alpha=0.05) were combined in a second model (fourth and fifth columns). This model included all of the above covariates plus comparative body size aged 10, smoking status, age at menarche, and HLA genotype.

Table 3: additive interaction terms and 95% confidence intervals for significant categorical predictors of MS risk.

Table 1: demographic characteristics of included participants. Continuous variables are presented as mean(SD), categorical variables are presented as n(%). Missing data are not tabulated. Proportions are calculated as a proportion of individuals with non-missing data for each variable.

TRAIT	CONTROLS	CASES
AGE	56.54 (8.09)	55.16 (7.64)
SEX		
FEMALE	263150 (54.13 %)	1563 (72.66 %)
MALE	222975 (45.87 %)	588 (27.34 %)
COUNTRY OF BIRTH		
UK	446458 (92.09 %)	2060 (95.99 %)
NON-UK	38324 (7.91 %)	86 (4.01 %)
ETHNICITY		
WHITE	458046 (94.69 %)	2100 (98.22 %)
NON-WHITE	25669 (5.31 %)	38 (1.78 %)
A*02 ALLELES		
0	264819 (54.48 %)	1366 (63.51 %)
1	186045 (38.27 %)	676 (31.43 %)
2	35261 (7.25 %)	109 (5.07 %)
DRB1*15 ALLELES		
0	360497 (74.16 %)	1084 (50.4 %)
1	115808 (23.82 %)	914 (42.49 %)
2	9820 (2.02 %)	153 (7.11 %)
BIRTH LATITUDE	360105.64 (162180.95)	359723.28 (167494.39)
BIRTH WEIGHT (KG)	3.32 (0.67)	3.28 (0.68)
MONTH OF BIRTH		
APRIL	41724 (8.58 %)	184 (8.55 %)
AUGUST	40074 (8.24 %)	186 (8.65 %)
DECEMBER	39050 (8.03 %)	161 (7.48 %)
FEBRUARY	38684 (7.96 %)	168 (7.81 %)
JANUARY	41058 (8.45 %)	172 (8 %)
JULY	41201 (8.48 %)	181 (8.41 %)
JUNE	40989 (8.43 %)	176 (8.18 %)
MARCH	43665 (8.98 %)	194 (9.02 %)
MAY	43666 (8.98 %)	201 (9.34 %)
NOVEMBER	37129 (7.64 %)	174 (8.09 %)
OCTOBER	39266 (8.08 %)	183 (8.51 %)

TRAIT	CONTROLS	CASES
SEPTEMBER	39619 (8.15 %)	171 (7.95 %)
BREASTFED		
NO	102542 (27.61 %)	537 (30.72 %)
YES	268847 (72.39 %)	1211 (69.28 %)
MATERNAL SMOKING		
NO	296365 (70.73 %)	1311 (70.67 %)
YES	122652 (29.27 %)	544 (29.33 %)
AGE COMPLETED FULL-TIME EDUCATION	16.72 (2.33)	16.95 (2.41)
AGE HAD SEXUAL INTERCOURSE	19.11 (3.89)	18.75 (3.87)
AGE AT MENARCHE	12.97 (1.62)	12.78 (1.66)
COMPARATIVE BODY SIZE AGED 10		
THINNER	158648 (42.72 %)	581 (33.24 %)
AVERAGE	241829 (65.11 %)	1104 (63.16 %)
PLUMPER	75381 (20.3 %)	427 (24.43 %)
AGE AT VOICE BREAKING		
AVERAGE	182877 (89.71 %)	481 (87.77 %)
YOUNGER	8926 (4.38 %)	33 (6.02 %)
OLDER	12043 (5.91 %)	34 (6.2 %)
SMOKER AGED <20		
NO	394287 (81.11 %)	1718 (79.87 %)
YES	91838 (18.89 %)	433 (20.13 %)
PREVIOUS IM		
NO	484970 (99.76 %)	2139 (99.44 %)
YES	1155 (0.24 %)	12 (0.56 %)
CURRENT TOWNSEND DEPRIVATION INDEX	-1.31 (3.09)	-1.38 (3.06)

Table 2: odds ratios for Multiple Sclerosis for each exposure studied. The first three columns depict the multivariable odds ratios from the output of regression models incorporating age, sex, ethnicity, birth latitude and current deprivation as covariates. Predictors which conferred good model fit (likelihood ratio p value < multiple testing threshold for alpha=0.05) were combined in a second model (fourth and fifth columns). This model included all of the above covariates plus comparative body size aged 10, smoking status, age at menarche, and HLA genotype.

	MODEL ADJUSTED FOR AGE, SEX, ETHNICITY, BIRTH LATITUDE, AND DEPRIVATION			MODEL ADJUSTED FOR AGE, SEX, ETHNICITY, BIRTH LATITUDE, DEPRIVATION, AND INCLUDING ALL COVARIATES ASSOCIATED WITH MS	
	OR (95% CI)	**P Value**	**Likelihood ratio p value**	**OR (95% CI)**	**P Value**
HLA DRB1*15 ALLELES			3.04E-115		
1	2.55 (2.33 - 2.8)	1.69E-88		2.54 (2.31 - 2.79)	2.48E-86
2	5.04 (4.22 - 6.02)	4.92E-71		5.05 (4.22 - 6.05)	3.46E-70
HLA A*02 ALLELES			1.36E-20		
1	0.66 (0.6 - 0.73)	6.13E-17		0.66 (0.6 - 0.73)	1.86E-16
2	0.57 (0.46 - 0.69)	3.36E-08		0.57 (0.46 - 0.7)	5.79E-08
BIRTH WEIGHT	0.97 (0.89 - 1.06)	0.539171	0.539438985		
MONTH OF BIRTH			0.975713464		
AUG	1.04 (0.84 - 1.28)	0.732411			
DEC	0.93 (0.74 - 1.15)	0.499897			
FEB	0.95 (0.76 - 1.18)	0.649151			
JAN	0.93 (0.75 - 1.16)	0.518807			
JUL	0.96 (0.77 - 1.18)	0.678821			
JUN	0.94 (0.76 - 1.17)	0.577315			
MAR	1 (0.81 - 1.23)	0.994643			
MAY	1.01 (0.82 - 1.25)	0.8998			
NOV	1.03 (0.83 - 1.28)	0.778539			
OCT	1.07 (0.86 - 1.32)	0.553756			
SEPT	0.99 (0.8 - 1.23)	0.915099			
BREASTFED	0.99 (0.88 - 1.1)	0.798447	0.798604505		
COMPARATIVE BODY SIZE AGED 10			3.84E-06		
AVERAGE	1.19 (1.07 - 1.32)	0.001352		1.19 (1.07 - 1.32)	0.001385521
PLUMPER	1.39 (1.22 - 1.58)	7.67E-07		1.39 (1.22 - 1.58)	6.72E-07
EXPOSED TO MATERNAL SMOKING	0.94 (0.85 - 1.05)	0.275974	0.274170936		
AGE AT MENARCHE	0.93 (0.9 - 0.96)	4.54E-05	4.25E-05	0.95 (0.92 - 0.98)	0.002185041
RELATIVE AGE AT VOICE BREAKING			0.194365056		

YOUNGER THAN AVERAGE	1.41 (0.99 - 2.01)	0.059277			
OLDER THAN AVERAGE	0.98 (0.68 - 1.42)	0.927958			
SMOKER AGED <20	1.21 (1.08 - 1.35)	0.000911	0.001124345	1.19 (1.07 - 1.33)	0.002151235
AGE FIRST HAD SEXUAL INTERCOURSE	0.99 (0.97 - 1)	0.054173	0.05036614		
PREVIOUS IM	1.7 (0.88 - 3.28)	0.115743	0.147382119		

Table 3: additive interaction terms and 95% confidence intervals for significant predictors of MS risk and their interaction with HLA genotype and the genome-wide polygenic risk score (PRS).

ENVIRONMENTAL EXPOSURE	GENETIC RISK FACTOR	AP	95% LOWER CI	95% UPPER CI	P VALUE
CHILDHOOD BMI	A*02	-0.150760249	-0.45315	0.100878	0.250575
MENARCHE	A*02	0.114768285	-0.04547	0.491886	0.233177
SMOKING AGED <20	A*02	0.042817986	-0.21541	0.255306	0.738926
CHILDHOOD BMI	DRB1*15	0.086845636	-0.06554	0.204275	0.234377
MENARCHE	DRB1*15	-0.041392527	-0.10011	0.042104	0.271773
SMOKING AGED <20	DRB1*15	0.007479754	-0.15018	0.134183	0.924508
CHILDHOOD BMI	PRS	0.111913013	0.007967	0.202299	0.035996
MENARCHE	PRS	-0.018559382	-0.07235	0.051991	0.579342
SMOKING AGED <20	PRS	0.097967464	-0.01324	0.194468	0.081792

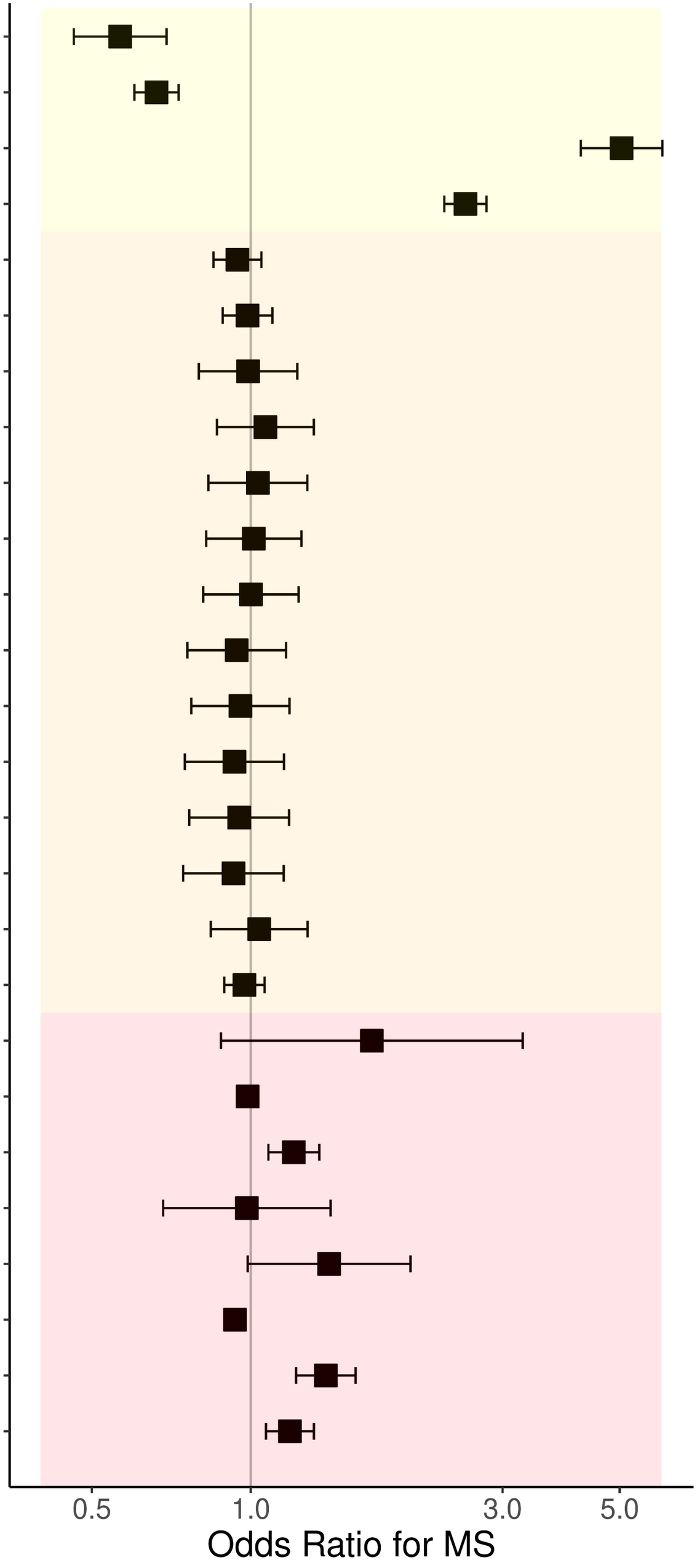
Supplementary table 1: definitions of covariates extracted from UK Biobank.

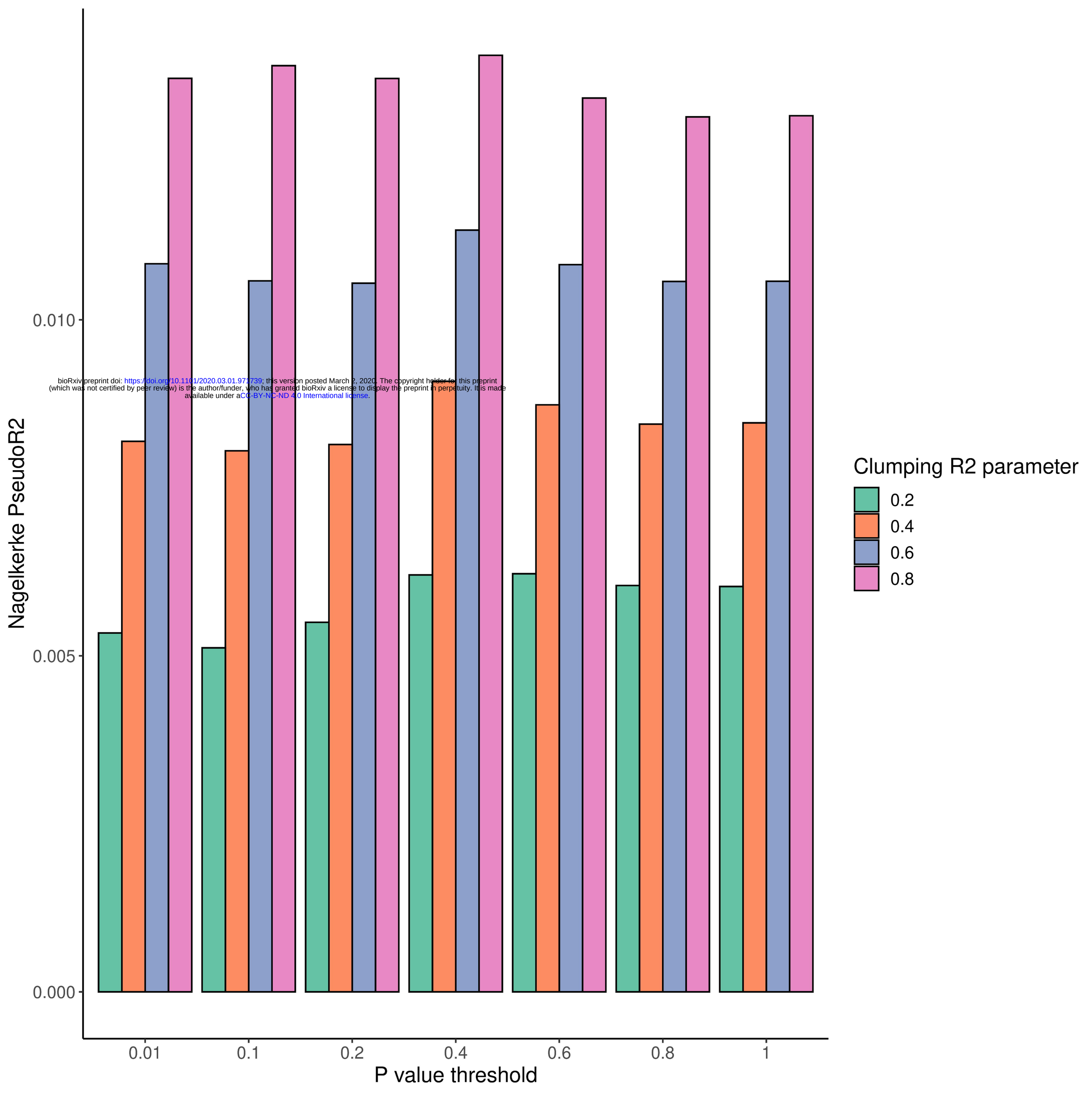
Variable	Method of data acquisition	UK Biobank URL
Sex	Registry	http://biobank.ndph.ox.ac.uk/showcase/field.cgi?id=31
Month of birth	Registry	http://biobank.ndph.ox.ac.uk/showcase/field.cgi?id=52
Age at recruitment	Derived from birth date and date of 1st assessment	http://biobank.ndph.ox.ac.uk/showcase/field.cgi?id=21022
Place of Birth in UK North Co-ordinate (latitude)	Verbal interview	http://biobank.ndph.ox.ac.uk/showcase/field.cgi?id=129
Townsend Deprivation Index at recruitment	Derived from census data and postcode	http://biobank.ndph.ox.ac.uk/showcase/field.cgi?id=189
Breastfeeding status	Touchscreen question: “Were you breastfed when you were a baby?”	http://biobank.ndph.ox.ac.uk/showcase/field.cgi?id=1677
Age at voice breaking	Touchscreen question "When did your voice break?"	http://biobank.ndph.ox.ac.uk/showcase/field.cgi?id=2385
Age at menarche	Touchscreen question "How old were you when your periods started?"	http://biobank.ndph.ox.ac.uk/showcase/field.cgi?id=2714
Age first had sexual intercourse	Touchscreen question "What was your age when you first had sexual intercourse? (Sexual intercourse includes vaginal, oral or anal intercourse)"	http://biobank.ndph.ox.ac.uk/showcase/field.cgi?id=2139
Smoking status	Derived from 2 touchscreen questions: 1. "Do you smoke tobacco now?" 2. "In the past, how often have you smoked	http://biobank.ndph.ox.ac.uk/showcase/field.cgi?id=20116 http://biobank.ndph.ox.ac.uk/showcase/field.cgi?id=3436 http://biobank.ndph.ox.ac.uk/showcase/field.cgi?id=2867

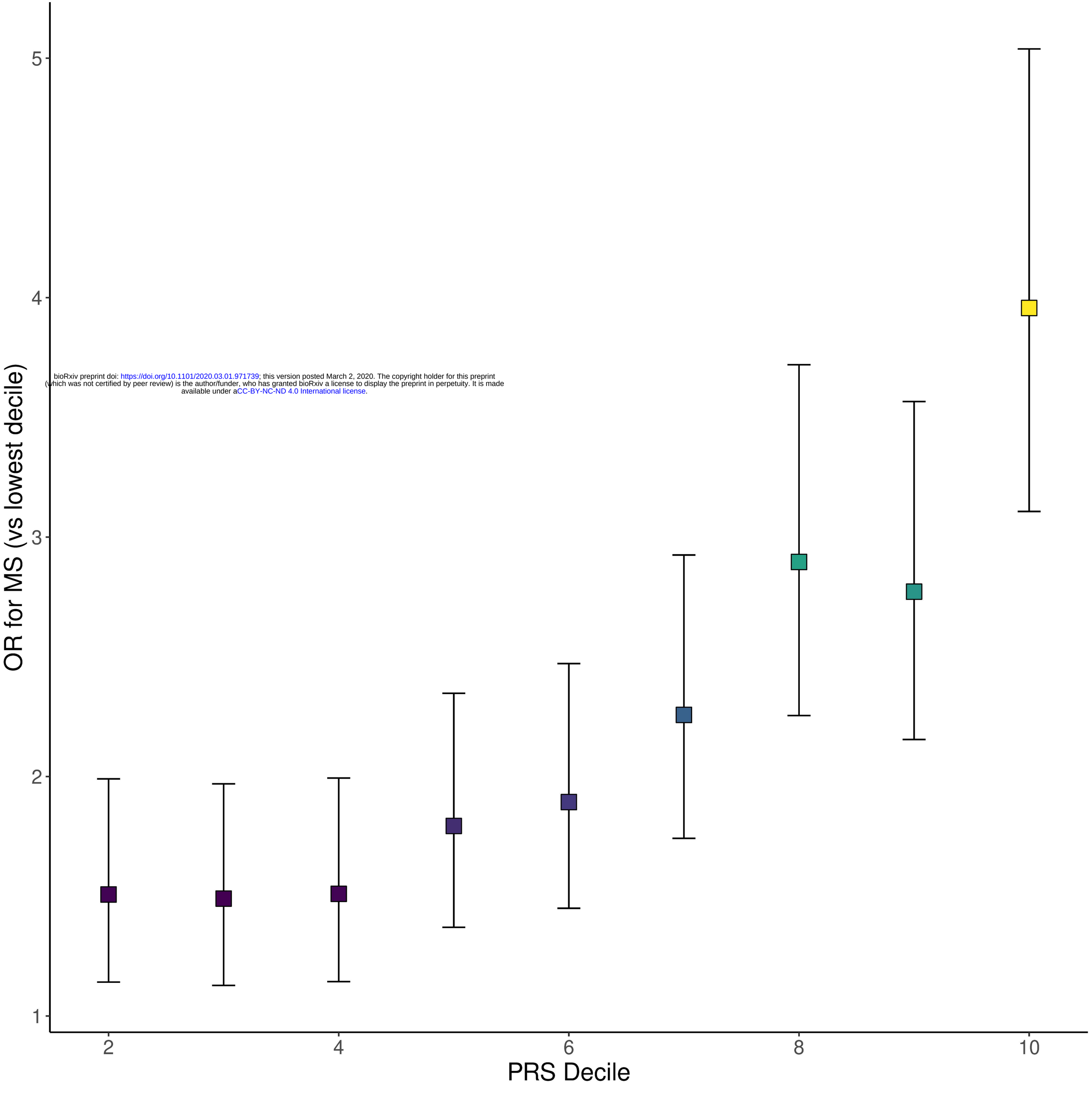
	tobacco?" And for current/previous smokers, age when started smoking was derived from the question "How old were you when you first started smoking on most days?"	
Comparative body size aged 10	Touchscreen question: "When you were 10 years old, compared to average would you describe yourself as:"	http://biobank.ndph.ox.ac.uk/showcase/field.cgi?id=1687
Ethnicity	Derived from several touchscreen questions	http://biobank.ndph.ox.ac.uk/showcase/field.cgi?id=21000
Birth weight (Kg)	Touchscreen	http://biobank.ndph.ox.ac.uk/showcase/field.cgi?id=20022
Infectious mononucleosis	Self-reported	http://biobank.ndph.ox.ac.uk/showcase/field.cgi?id=20002
Multiple sclerosis	HES records - ICD10 G35; ICD9 3409 Self-reported	Multiple
Maternal smoking around the time of birth	Touchscreen question: "Did your mother smoke regularly around the time when you were born?"	http://biobank.ndph.ox.ac.uk/showcase/field.cgi?id=1787
Other diseases (PD, SLE, RA, IBD etc)	Self-reported	http://biobank.ndph.ox.ac.uk/showcase/field.cgi?id=20002

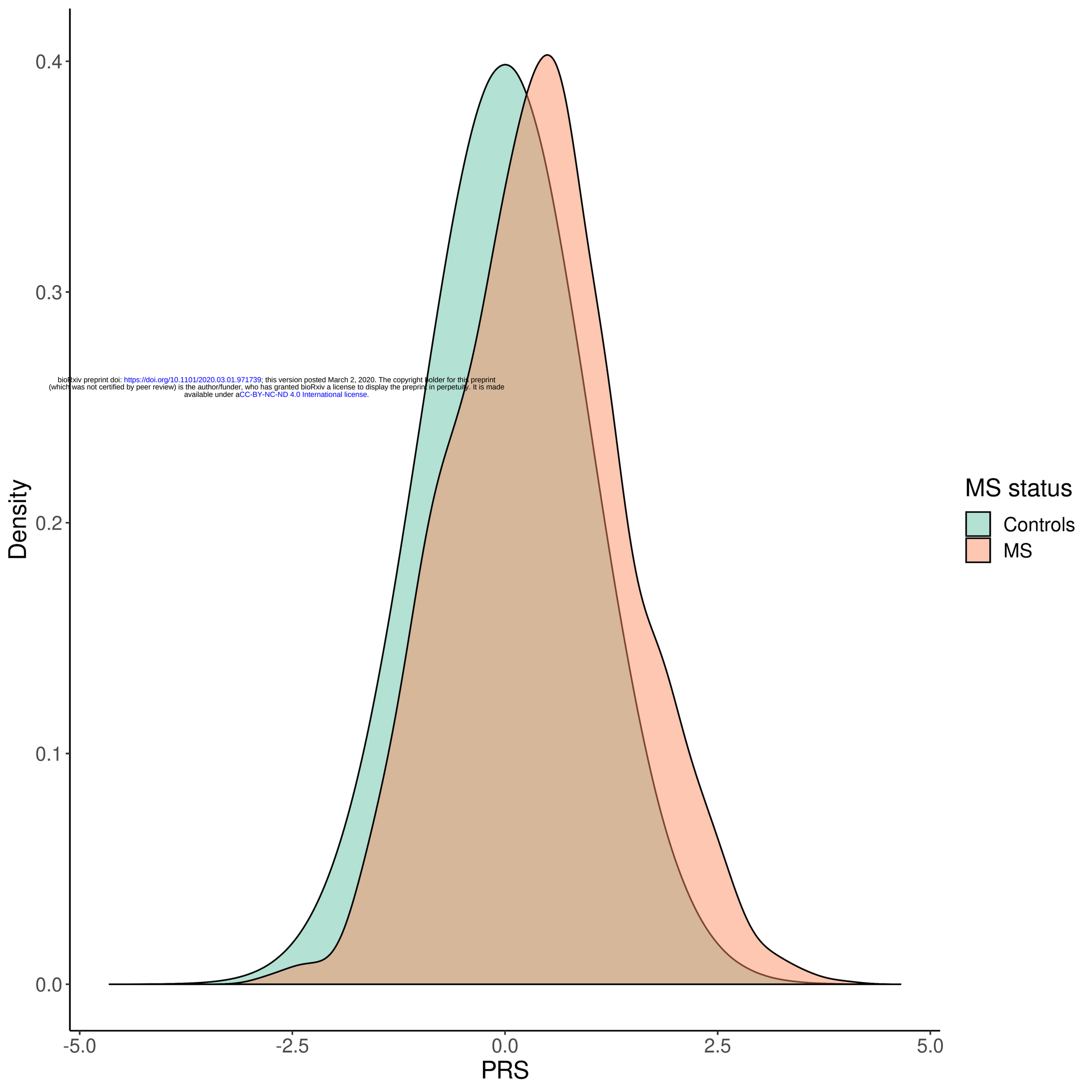
Risk factor

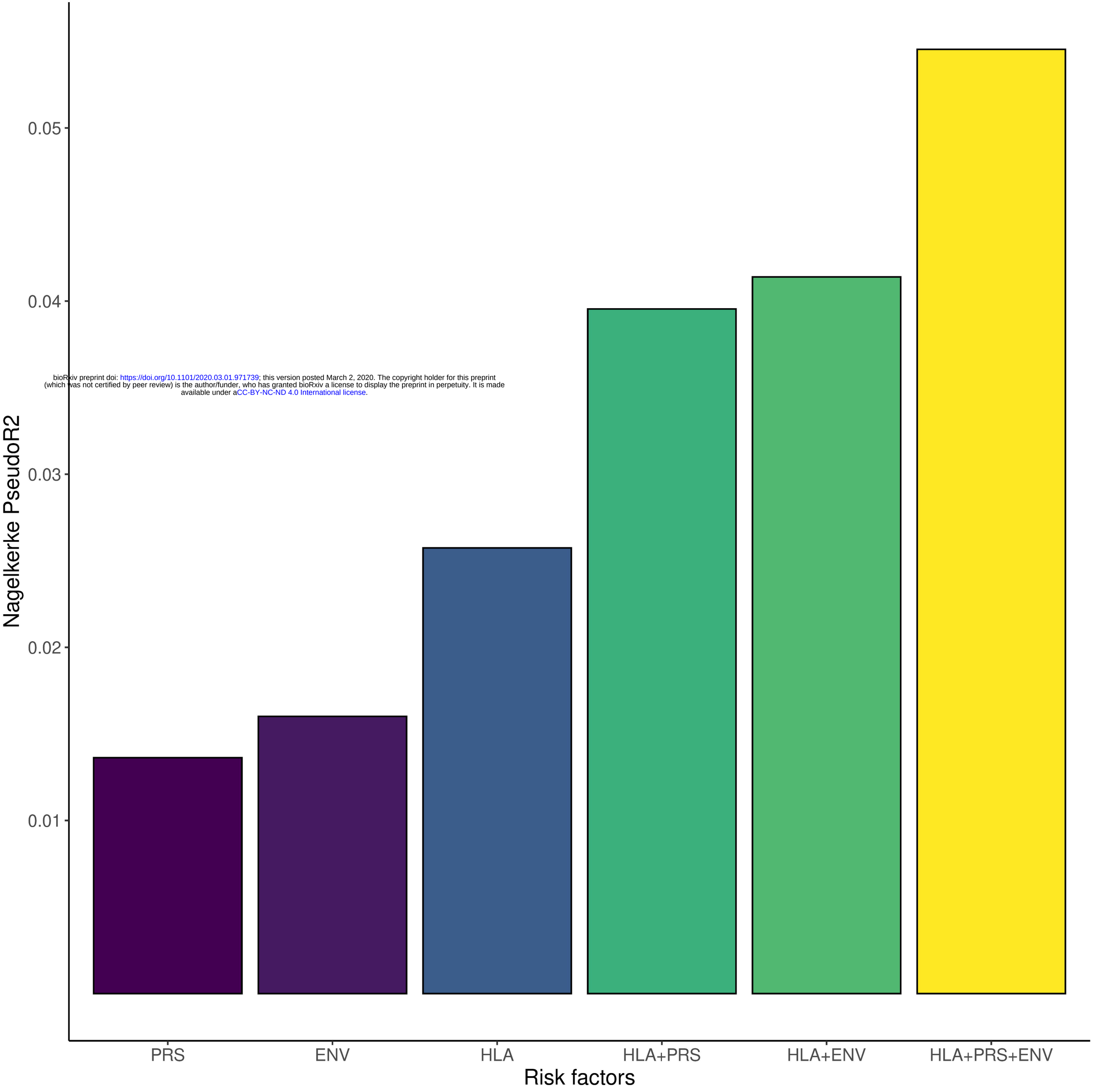
HLA A*02 homozygote (vs no A*02 alleles)
HLA A*02 heterozygote (vs no A*02 alleles)
HLA DRB1*15 homozygote (vs no DRB1*15 alleles)
HLA DRB1*15 heterozygote (vs no DRB1*15 alleles)
Exposed to maternal smoking (vs not)
Breastfed as a baby (vs not)
Born in September (vs April)
Born in October (vs April)
Born in November (vs April)
Born in May (vs April)
Born in March (vs April)
Born in June (vs April)
Born in July (vs April)
Born in January (vs April)
Born in February (vs April)
Born in December (vs April)
Born in August (vs April)
Birth weight
Ever had IM
Age of first sexual intercourse
Smoker aged <20
Age when voice broke (males): older than average (vs average)
Age when voice broke (males): younger than average (vs average)
Age at menarche (females)
Comparative body size aged 10: plumper (vs thinner)
Comparative body size aged 10: average (vs thinner)

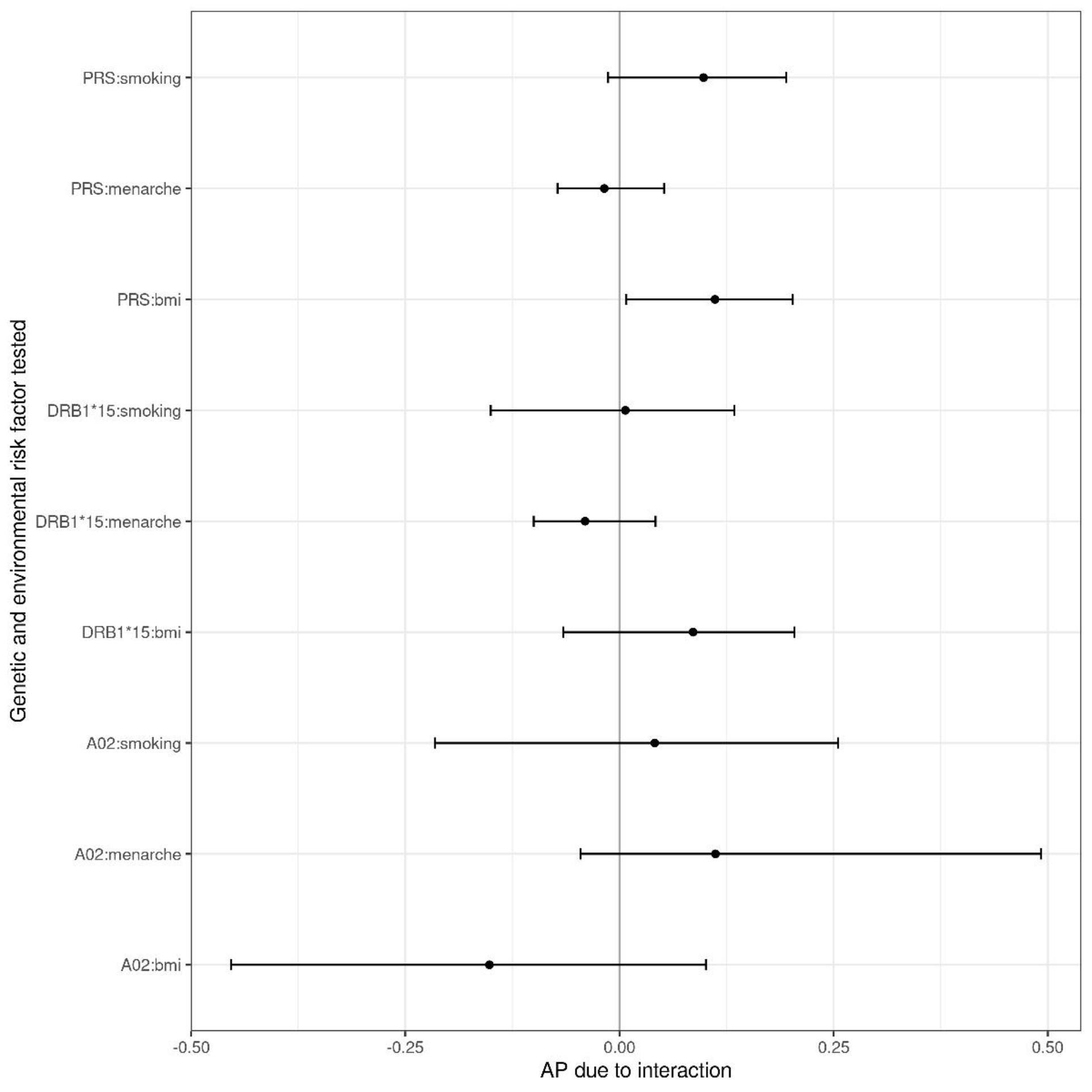











PRS:smoking
PRS:menarche
PRS:bmi
DRB1*15:smoking
DRB1*15:menarche
DRB1*15:bmi
A02:smoking
A02:menarche
A02:bmi
Genetic and environmental risk factor tested
-0.50
-0.25
0.00
0.25
0.50
AP due to interaction