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 (MBBS) [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.

Abstract word count: 272

Manuscript word count: 3233

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 available 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 neuro-inflammatory disorder caused by a combination of environmental exposures and genetic risk factors. We sought to determine whether genetic risk modifies the effect of environmental MS risk factors.

Methods:
People with MS were identified within UK Biobank using ICD10-coded MS or self-report. Associations between environmental risk factors and MS risk were quantified with a case-control design using multivariable logistic regression. Polygenic risk scores (PRS) were derived using the clumping-and-thresholding approach with external weights from the largest genome-wide association study of MS. Separate scores were created including (PRS_{MHC}) and excluding ($PRS_{Non\text{-}MHC}$) the MHC locus. The best performing PRS were identified in 30% of the cohort and validated in the remaining 70%. Interaction between environmental and genetic risk factors was quantified using the Attributable Proportion due to interaction (AP) and multiplicative interaction.

Results:
Data were available for 2250 people with MS and 486,000 controls. Childhood obesity, earlier age at menarche, and smoking were associated with MS. The optimal PRS were strongly associated with MS in the validation cohort (PRS_{MHC}: Nagelkerke's Pseudo-R^2 0.033, $p=3.92x10^{-111}$; $PRS_{Non\text{-}MHC}$: Nagelkerke's Pseudo-R^2 0.013, $p=3.73x10^{-43}$). There was strong evidence of interaction between polygenic risk for MS and childhood obesity (PRS_{MHC}: AP=0.17, 95% CI 0.06 - 0.25, p=0.004; $PRS_{Non\text{-}MHC}$: AP=0.17, 95% CI 0.06 - 0.27, p=0.006).

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

Introduction

Susceptibility to Multiple Sclerosis (MS) is multifactorial with genetic and environmental determinants. [1,2,3]. Environmental exposures associated with MS risk include smoking, solvent exposure, childhood obesity, vitamin D deficiency, increasing latitude, and infectious mononucleosis (IM) [2,3]. The largest genome-wide association study (GWAS) of MS risk performed by the International Multiple Sclerosis Genetic Consortium (IMSGC) revealed 233 independent signals which account for ~48% of the estimated heritability of MS[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 common variants across the genome, in addition to the contribution from the Major Histocompatibility Complex (MHC).

A large proportion of MS risk remains unexplained despite the well-described genetic architecture[1]. One potential explanation for this 'missing risk' may is the presence of gene-environment interactions, whereby the effect of certain genes or variants may depend on exposure to environmental risk factors.

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, IM, and solvent exposure on MS risk are potentiated among carriers of the HLA DRB1*15 allele and those lacking the protective HLA A*02 genotype[7–10]. It is not currently known whether gene-environment interactions in MS extend beyond the HLA locus[11,12].

In this work, we harnessed the power of UK Biobank to extend our understanding of how common genetic variation interacts with environmental factors associated with MS development. We achieved this by firstly performing a large case-control study to confirm the role of established risk factors in this cohort, and by developing and validating polygenic risk scores (PRS) for MS which both included and excluded the MHC. Finally, we used these data to look for potential interactions between polygenic risk and environmental factors associated with MS.

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. Phenotype data are composed of survey data, linked healthcare records, anthropometric measurements, and a variety of other biochemical and imaging data (which was not used in this study).

Identification of cases and controls

Cases were defined by ICD-coded diagnoses (ICD10-G35; ICD9-3409), self-reported MS diagnosis, or a GP-coded diagnosis. Age at diagnosis was determined using the first recorded MS diagnostic code (see supplementary methods for further details). Controls were unmatched UK Biobank participants without a coded diagnosis of MS. Individuals diagnosed with MS prior to age 20 were excluded due to concerns around the timing of exposures. Participant flow through the study is depicted in supplementary figure 1; diagnostic codes used are provided in supplementary data.

Genotype data

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 panel (see https://biobank.ctsu.ox.ac.uk/crystal/crystal/docs/HLA_imputation.pdf). 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]**.** Genetic principal components were supplied by UK Biobank (field ID 22009).

Definition of exposures

Exposures were selected if they pertained to early life/adolescence (to mitigate the risk of reverse causation), and were previously associated with MS in at least one other observational cohort. Selected exposures were captured from baseline data recorded in UKB, along with age, ethnicity, sex, birth latitude, and Townsend deprivation index at recruitment (supplementary table 1).

We examined the following ten early life/environmental exposures: month of birth, having been breastfed as a child, childhood body size at age 10 (CBS_{10}, a proxy for childhood obesity), exposure to maternal smoking, age at menarche (females), age at voice breaking (males), age at first sexual intercourse, smoking status prior to age 20, birth weight, and infectious mononucleosis prior to age 20. Where multiple data points were available for a participant, the first recorded reading was used.

Childhood body size was dichotomised and 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. Age at menarche was treated as a continuous variable and analyses regarding menarche were restricted to women. IM status prior to age 20 was defined using the source of first report fields. Participants whose IM diagnosis was reported after age 20 were coded as having not had IM. Vitamin D status was not included, as vitamin D levels are only available from the initial visit (i.e. at study recruitment), which in the majority of cases was subsequent to diagnosis.

Case-control study

For each risk factor, we built a multivariable logistic regression model modelling MS status as the outcome, with age, sex, ethnicity, current deprivation status, and birth latitude as potential confounding covariates[15].

The strength of evidence for association with MS was determining using the model likelihood ratio, comparing the full model to a null model comprising only the confounding covariates. Strong evidence for association was defined using a Bonferroni-adjusted p-value threshold to maintain an alpha of 0.05 ($p_{threshold}$=0.05/10=0.005). Risk factors robustly associated with MS at alpha<0.05 were then combined in a multivariable model including the most potent genetic risk factors, HLA DRB1*15:01 and HLA A*02:01, to assess whether their effects showed evidence of independent association with MS.

Development of polygenic risk scores (PRS) for MS

A variety of PRS were created using the clumping-and-thresholding approach with external weights derived from the IMSGC discovery stage meta-analysis (supplementary methods).

We created scores both including the MHC region (PRS_{MHC}) and excluding this region ($PRS_{Non\text{-}MHC}$). To validate the PRS, the dataset was divided randomly into a training set (30%, n_{MS}=589, $n_{control}$=112,724) and a testing set (70%, n_{MS}=1237, $n_{control}$=263,159, supplementary figure 1). To determine the optimal PRS, we constructed multivariable logistic regression models for each PRS with MS status as the outcome with age, sex, Townsend deprivation index, and the first four genetic principal components (PCs) as confounding covariates.

PRS performance was evaluated using Nagelkerke's Pseudo-R^2 metric, which is analogous to the R^2 derived from linear regression models. Nagelkerke's Pseudo-R^2 was calculated comparing the full model including the PRS to a null model comprising the confounding covariates only. This procedure was repeated for all 64 scores (supplementary table 4). Altering the number of PCs adjusted for did not substantially alter the results (supplementary figures 5&6). Further validation is described in the supplementary methods.

PRS x Environment interactions

The optimal PRS_{MHC} and $PRS_{Non\text{-}MHC}$ were used to look for evidence of genome-wide gene-environment interactions using exposures identified as significantly associated with MS in the case-control study. All interaction analyses were conducted in the testing set to avoid PRS overfitting. Interaction was assessed on the additive and multiplicative scales (supplementary methods for full details). Multiplicative interaction was quantified using the interaction term beta from logistic models, and additive interaction was quantified using the Attributable Proportion due to interaction (AP).

HLA x PRS interactions

To determine whether non-MHC genetic risk of MS modulates the effects of the most potent MHC risk allele, DRB1*15:01, we calculated additive and multiplicative interaction statistics using the methods described above, considering both the DRB 1*15:01 genotype (dominant-coding) and the non-MHC PRS as independent covariates.

Association of PRS with disease measures

To determine whether the MS-PRS was associated with age at first report and disability status, we constructed regression models in the testing set. For age at first MS diagnostic code report, values were normalised using the inverse-rank normalisation. Linear regression models were constructed, using age, sex, Townsend score, and the first four PCs as covariates. Disability status was assessed using the UKB field 'Attendance/disability/mobility allowance' (field 6146), and recoded this as a binary variable (i.e. participants were coded as '1' if they claimed any of the blue badge, attendance allowance, or disability living allowance, and as '0' if not). Logistic regression models were then constructed using the same covariates as above (age, sex, Townsend score, first four genetic PCs).

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[19]. Statistical analyses were performed in R version 3.6.1.

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. All code used in this study is available on GitHub (@benjacobs123456).

Results

Population demographics

Phenotype and genotype data were available for 488,276 UK Biobank participants comprising 2276 people with MS and 486,000 unmatched controls. The median age at first MS report was 43.5 (IQR 16.1, supplementary figure 2). Demographic characteristics are shown in table 1. Characteristics of individuals with MS were consistent with published observational data (72.7% female, 98.1% White British).

Exposures associated with MS in UK Biobank

There was strong evidence for association between three of the ten risk factors examined and MS ($P_{Bonf}<0.05$): higher childhood body size at age 10 ('plumper than average' vs 'thinner than average': OR 1.36, 95% CI 1.20 - 1.55), smoking prior to age 20 (OR 1.21, 95% CI 1.08 - 1.34), and earlier menarche (OR 0.94, 95% CI 0.91 - 0.97, figure 1, table 1). The effects of these three risk factors remained similar in a combined model incorporating HLA DRB1*15:01 and HLA A*02:01 genotype (supplementary table 3).

Development and validation of PRS for MS

The optimal PRS_{MHC} and $PRS_{Non\text{-}MHC}$ explained 3.5% and 1.3% of MS risk in the training set respectively (figure 2a, supplementary table 4, supplementary figures 5 and 6). Both scores were strongly associated with MS in the testing set (PRS_{MHC}: Nagelkerke's Pseudo-R^2 0.033, $p=3.92x10^{-111}$; $PRS_{Non\text{-}MHC}$: Nagelkerke's Pseudo-R^2 0.013, $p=3.73x10^{-43}$, figure 2, supplementary table 5). Both scores were reasonably well-calibrated (figure 3a) with good discriminative performance (AUC_{MHC} 0.71, $AUC_{non\text{-}MHC}$ 0.67, AUC_{null} 0.63; figure 3b). There was no evidence of association between the PRS_{MHC} and $PRS_{Non\text{-}MHC}$ and age at MS report (figure 3c & 3d) or claiming of disability benefits ($p_{MHC}=0.44$, $p_{Non\text{-}MHC}=0.96$, supplementary figure 7).

*PRS interactions with environmental risk factors and DRB1*15:01*

We found strong evidence of interaction on the additive scale between the PRS_{MHC} and $PRS_{Non\text{-}MHC}$ and childhood body size (PRS_{MHC}: AP=0.17, 95% CI 0.06 - 0.25, p=0.004; $PRS_{Non\text{-}MHC}$: AP=0.17, 95% CI 0.06 - 0.27, p=0.006). We found weaker evidence for interaction on this scale between age at menarche and the PRS_{MHC} (AP=-0.05, 95% CI -0.10 - 0.00, p = 0.033; figure 4a, table 2), but this estimate did not surpass the multiple testing threshold (table 2). There was a lack of strong evidence for other pairwise additive interactions (figure 4) or for multiplicative interactions (supplementary figure 8, supplementary table 6). There was evidence of additive interaction between the $PRS_{Non\text{-}MHC}$ and HLA DRB1*15:01 carriage (AP 0.24, 95% CI 0.17 - 0.30, p=0.0002, figure 4b), but no evidence of multiplicative interaction (beta 0.060, p=0.30).

Discussion

In this study we harnessed the scale and breadth of UK Biobank to study >2000 MS cases and >480,000 controls, providing the first evidence that the effect of an established risk factor

for MS (childhood obesity) may be potentiated by an individual's genome-wide genetic risk for MS. We show that this effect persists even when the MHC locus is excluded from the PRS. By using data from the largest GWAS of MS susceptibility to derive and validate polygenic risk scores for MS, both incorporating and excluding the MHC region, we demonstrate supportive evidence for a gene-gene interaction. This work shows that the effect of DRB1*15:01 on MS susceptibility may be potentiated among individuals with a high background genetic risk for MS in this cohort. To our knowledge, our study is the first to demonstrate that the polygenic risk of an individual for MS may alter the effect of established environmental risk factors on their risk of MS[3]. These findings are especially interesting in the context of evidence from mendelian randomisation studies supporting a causal role for childhood obesity in the pathogenesis of MS[21,22].

Previous studies of gene-environment interactions in MS have focussed on interactions between HLA alleles and environmental risk factors. Specifically, evidence suggests that carriage of high-risk HLA haplotypes containing DRB1*15:01 and lacking A*02:01 enhances the deleterious association of childhood obesity, smoking, Infectious Mononucleosis, and solvent exposure with MS risk[2,789]. The intuitive biological explanation for such interactions is that high-risk HLA alleles may promote presentation of epitopes, e.g. from cigarette smoke or within adipose tissue, in such a way that mimics myelin peptides and triggers CNS-directed autoimmunity. Beyond the MHC, there has been relatively limited study of how genetic variation modulates the effect of environmental risk factors for MS[11,12], probably in large part due to the relatively small number of datasets with sufficient power, deep phenotyping, and high-quality genetic data required for such analyses.

In this study we created 64 individual PRS, both including and excluding the MHC locus on chromosome 6, which is the strongest single genetic determinant of MS risk and accounts for ~20% of the SNP heritability of MS in Europeans[1]. Both the non-MHC and MHC PRS were strongly associated with MS risk in both training and testing sets. The non-MHC PRS in this study 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 variance[5]. The best-performing non-MHC PRS in this study explained ~1% of MS variance. This discrepancy could be explained by several factors, including the relatively low number of cases in UK Biobank, the possibility of missed cases, differences in population structure, restriction according to self-declared ethnicity with an additional genetic principal components analysis, and some SNPs not being available and/or failing QC checks in Biobank. Nevertheless, despite low overall variance, 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, reasonable discriminative capacity, and good calibration.

There are several important caveats to this work. Most importantly, while we are able to observe and measure statistical interaction – i.e. deviation from a model whereby the effects of genetic and environmental risk factors are combined additively (in the case of the AP) or multiplicatively (in the case of multiplicative interaction – statistical interaction does not straightforwardly imply biological interaction, nor does it necessarily imply interaction that is meaningful in terms of real-life disease prediction or prevention. We were unable to demonstrate replication in a truly independent cohort (dividing the cohort into training and testing sets does not yield a genuinely independent cohort). Our findings have limited generalisability for non-European groups as UK Biobank participants are predominantly White. MS diagnosis in this cohort are derived from linked healthcare records or self-report, so do not carry the same degree of certainty as criteria-defined MS. Equally, it is conceivable

that there are 'missed' cases in the dataset, i.e. individuals with MS who do not have a coded diagnosis available through linked healthcare records. However, MS prevalence in UK Biobank approaches the expected UK prevalence, suggesting that the overwhelming majority of individuals with MS are correctly identified. The UKB cohort is highly selected, and is enriched for individuals living near assessment centres, from more affluent socio-economic groups than the general population, for White British individuals, and (intentionally) for individuals over the age of 40 (the minimum age at recruitment). These factors carry a risk of introducing various biases, e.g. through collider bias, which may induce spurious associations and destroy true associations. Our findings concerning gene-gene interactions could be replicated in 'genetics-only' cohorts such as the IMSGC, and we would encourage others to attempt to replicate this finding in large GWAS cohorts (with many more cases than the ~2000 in UKB) so we can ascertain whether it is robust.

The key variables used in this study are retrospective or cross-sectional (e.g. MS diagnosis, self-reported body size in childhood, self-reported smoking status). Not only are these subject to recall bias, but more importantly our results are not revealing about predicting an individual's risk of developing MS. To demonstrate *predictive* power, these results need to be replicated in a longitudinal cohort. In addition, the metric we focus on, 'comparative body size at age 10', is clearly not a perfect proxy for childhood obesity. Other limitations to this study include the limited overall variance explained by both optimal PRS, the relatively small absolute number of people with MS, and the imperfect nature of self-reported phenotypes. Furthermore, some exposures known to be strongly associated with MS were either unavailable (e.g. vitamin D status prior to diagnosis) or so under-reported as to be unreliable (e.g. Infectious Mononucleosis).

Despite these limitations, our study also has some strengths. We use the UK Biobank dataset, which provides a unique opportunity to study gene-environment interactions on a large-scale. The vast number of controls in UKB adds substantial power. We tune and test the PRS in separate samples, which is important to prevent overfitting of the PRS to the data. We use an agnostic approach to develop the PRS, using a range of clumping-and-thresholding parameters to discover the optimal structure of the PRS, allowing us to discover a significant improvement in predictive power from using a large number of variants weakly associated with MS over using strictly 'GWAS-significant' hits ($p<5e-8$). These optimal parameters also reiterate the polygenic architecture of MS.

We evaluate interactions on both the multiplicative and additive scales, as has become standard practice to avoid missing biologically-significant interactions[2]. We additionally evaluate the relationship between the PRS and proxies for clinical characteristics of MS, including age at diagnosis and disability status. We evaluated whether the effect of DRB1*15:01 is modulated by polygenic risk, as has been demonstrated for high-effect variants in the LDL-R (causing Familial Hypercholesterolaemia) and BRCA (causing breast cancer)[23], and find evidence in support of this hypothesis. Clearly, this finding is easily replicated in the IMSGC cohort and we would urge caution in overinterpreting the finding without confirmation in this far larger cohort of cases.

This study thus provides novel evidence that childhood body size interacts with non-HLA MS genetic risk. 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 the greatest benefit from the intervention. As the

effect of childhood body size on MS risk appears greater among individuals with a high genome-wide genetic risk, trials attempting to demonstrate benefit of targeting childhood obesity may benefit from risk-stratifying individuals using this approach. Further efforts are required to localise the variants and genes which account for the observed interaction effects, which should help to shed further light on the biology of these risk factors and improve efforts to individualise MS risk prediction algorithms in the future.

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. Briggs, F. B. S. *et al.* Smoking and risk of multiple sclerosis: evidence of modification by NAT1 variants. *Epidemiology* **25**, 605–614 (2014).
12. Briggs, F. B. Nicotinic acetylcholine receptors α7 and α9 modifies tobacco smoke risk for multiple sclerosis. *Mult. Scler.* 1352458520958361 (2020).
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. Murray, S., Bashir, K., Penrice, G. & Womersley, S. J. Epidemiology of multiple sclerosis in Glasgow. *Scott. Med. J.* **49**, 100–104 (2004).
16. Consortium, T. 1000 G. P. & The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* vol. 526 68–74 (2015).
17. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
18. 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).
19. King, T., Butcher, S. & Zalewski, L. *Apocrita - High Performance Computing Cluster for Queen Mary University of London*. (2017). doi:10.5281/zenodo.438045.
20. Fahed, A. C. *et al.* Polygenic background modifies penetrance of monogenic variants conferring risk for coronary artery disease, breast cancer, or colorectal cancer. *Genetic and Genomic Medicine* (2019) doi:10.1101/19013086.
21. Mokry, L. E. *et al.* Obesity and Multiple Sclerosis: A Mendelian Randomization Study. *PLoS Med.* **13**, e1002053 (2016).

22. Jacobs, B. M., Noyce, A. J., Giovannoni, G. & Dobson, R. BMI and low vitamin D are causal factors for multiple sclerosis: A Mendelian Randomization study. *Neurol Neuroimmunol Neuroinflamm* **7**, (2020).
23. Fahed, A. C. *et al*. Polygenic background modifies penetrance of monogenic variants for tier 1 genomic conditions. *Nat. Commun.* **11**, 3635 (2020).

Table 1: demographic characteristics of included participants and results from the case-control study. 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. Column 4 shows odds ratios (ORs) and 95% confidence intervals (CIs) for Multiple Sclerosis for each exposure studied. Odds ratios represent the output of multivariable logistic regression models incorporating age, sex, ethnicity, birth latitude and current deprivation as covariates. Wald test P values represent the test of the null hypothesis beta = 0 for each term, whereas likelihood ratio P values represent the overall model fit compared to a null model comprising only confounding covariates. P values exceeding the Bonferroni multiple testing threshold (alpha=0.05) are shown in bold. Reference values for categorical covariates are denoted as 'REF'.

Trait	**Controls (N=486,000)**	**Cases (N=2250)**	**OR (95% CI)**	**Wald test P value**	**Likelihood ratio P value**
Sex					
Female	263058 (54.13%)	1635 (72.67%)			
Male	222942 (45.87%)	615 (27.33%)			
Age	56.54 (8.09)	55.17 (7.66)			
Birth latitude	360093.76 (162174.29)	361960.6 (168566.29)			
Age completed full-time education	16.72 (2.33)	16.96 (2.49)			
Townsend deprivation index	-1.31 (3.09)	-1.38 (3.06)			
Ethnic background					
White	457927 (94.69%)	2193 (98.08 %)			
Non white	25664 (5.31 %)	43 (1.92 %)			
HLA A*02:01 alleles					
0	264736 (54.47%)	1431 (63.6 %)			
1	186009 (38.27%)	704 (31.29 %)			
2	35255 (7.25 %)	115 (5.11 %)			
HLA DRB1*15:01 alleles					
0	360423 (74.16%)	1144 (50.84 %)			
1	115763 (23.82%)	948 (42.13 %)			
2	9814 (2.02 %)	158 (7.02 %)			
Country of birth					
UK	446343 (92.09%)	2151 (95.81 %)			
Non-UK	38314 (7.91 %)	94 (4.19 %)			
Age had	19.11 (3.89)	18.72 (3.81)	0.98 (0.97	0.015709	0.013768

Trait	Controls (N=486,000)	Cases (N=2250)	OR (95% CI)	Wald test P value	Likelihood ratio P value
sexual intercourse			to 1)		
Age at menarche	12.97 (1.62)	12.8 (1.66)	0.94 (0.91 to 0.97)	0.000116	**0.00011**
Birth weight (Kg)	3.32 (0.67)	3.28 (0.68)	0.98 (0.9 to 1.07)	0.603259	0.603452
Month of birth					0.931194
April	41716 (8.58 %)	188 (8.36 %)	REF	REF	
August	40064 (8.24 %)	194 (8.62 %)	1.06 (0.86 to 1.3)	0.610756	
December	39042 (8.03 %)	168 (7.47 %)	0.94 (0.76 to 1.17)	0.59689	
February	38673 (7.96 %)	178 (7.91 %)	0.98 (0.8 to 1.22)	0.888063	
January	41051 (8.45 %)	175 (7.78 %)	0.92 (0.75 to 1.14)	0.460529	
July	41190 (8.48 %)	190 (8.44 %)	0.97 (0.79 to 1.2)	0.812077	
June	40979 (8.43 %)	185 (8.22 %)	0.95 (0.77 to 1.18)	0.666742	
March	43654 (8.98 %)	203 (9.02 %)	1.02 (0.83 to 1.25)	0.883078	
May	43657 (8.98 %)	204 (9.07 %)	0.99 (0.81 to 1.22)	0.928091	
November	37124 (7.64 %)	178 (7.91 %)	1.03 (0.83 to 1.27)	0.800849	
October	39247 (8.08 %)	201 (8.93 %)	1.11 (0.9 to 1.36)	0.327942	
September	39603 (8.15 %)	186 (8.27 %)	1.04 (0.85 to 1.29)	0.681762	
Breastfed as a baby					0.731403
No	102506 (27.61 %)	565 (30.86 %)	REF	REF	
Yes	268781 (72.39 %)	1266 (69.14 %)	0.98 (0.88 to 1.09)	0.731136	
Comparative body size aged 10					**7.02E-06**
Thinner	158610 (42.72 %)	609 (33.26 %)	REF	REF	
About average	241759 (65.11 %)	1162 (63.46 %)	1.19 (1.08 to 1.32)	0.000697	
Plumper	75366 (20.3 %)	438 (23.92 %)	1.36 (1.2 to 1.55)	2.21E-06	

Trait	Controls (N=486,000)	Cases (N=2250)	OR (95% CI)	Wald test P value	Likelihood ratio P value
Exposed to maternal smoking					0.329681
No	296291 (70.73 %)	1368 (70.55 %)	REF	REF	
Yes	122618 (29.27 %)	571 (29.45 %)	0.95 (0.86 to 1.05)	0.331253	
Relative age at voicebreaking (males only)					0.132642
About average age	182848 (89.71 %)	504 (87.96 %)	REF	REF	
Younger than average	8924 (4.38 %)	35 (6.11 %)	1.44 (1.02 to 2.03)	0.038506	
Older than average	12043 (5.91 %)	34 (5.93 %)	0.95 (0.66 to 1.37)	0.776222	
Smoking status prior to age 20					**0.000915**
No	394188 (81.11 %)	1796 (79.82 %)	REF	REF	
Yes	91812 (18.89 %)	454 (20.18 %)	1.21 (1.08 to 1.34)	0.000737	
IM status					0.061221
No	484758 (99.75 %)	2235 (99.33 %)	REF	REF	
Yes	1238 (0.25 %)	15 (0.67 %)	1.82 (1.03 to 3.22)	0.040175	

Table 2: Attributable Propotion (AP) due to interaction, 95% CIs and two-sided p values for each of the PRS x E interactions examined. CIs represent the 2.5th and 97.5th centile from 10,000 bootstrap replicates. Two-sided p values represent absolute p values with a continuity correction, i.e. for a positive AP, the p value is given as: (number of iterations < 0 + 1)/(total number of iterations+1)*2

Interaction	AP	Lower CI	Upper CI	P value
MHC PRS x Childhood body size	0.167074	0.062196	0.254741	0.0042
Non-MHC PRS x Childhood body size	0.173705	0.055642	0.27455	0.005599
MHC PRS x Smoking	0.0768	-0.05055	0.177474	0.214179
Non-MHC PRS x Smoking	0.122975	-0.00556	0.228431	0.058794
MHC PRS x Age at menarche	-0.05206	-0.0968	-0.00478	0.033197
Non-MHC PRS x Age at menarche	0.021061	-0.04119	0.111064	0.551145

Figure and table legends

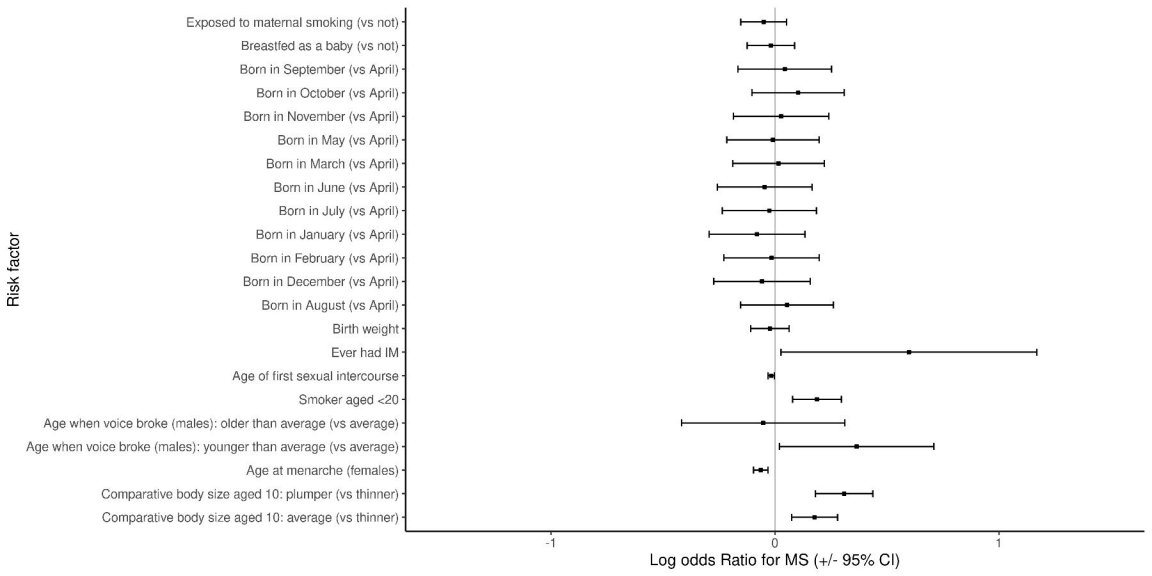
Figures

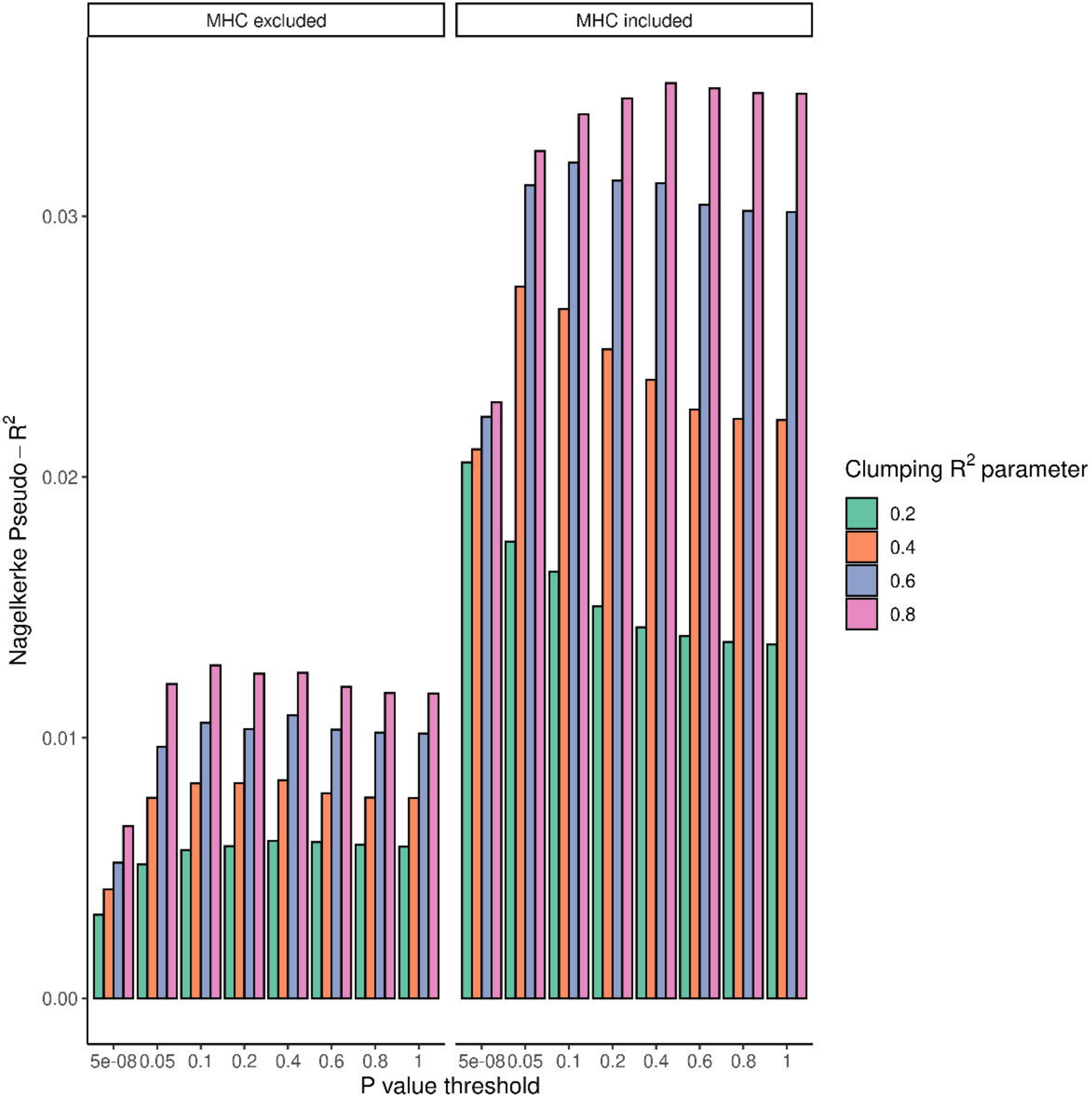
Figure 1: odds ratios and 95% confidence intervals for the association of each exposure with MS. ORs and CIs are from the output of a multivariable logistic regression with the following covariates: age, sex, ethnicity, birth latitude, current deprivation status, and the exposure in question. For menarche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

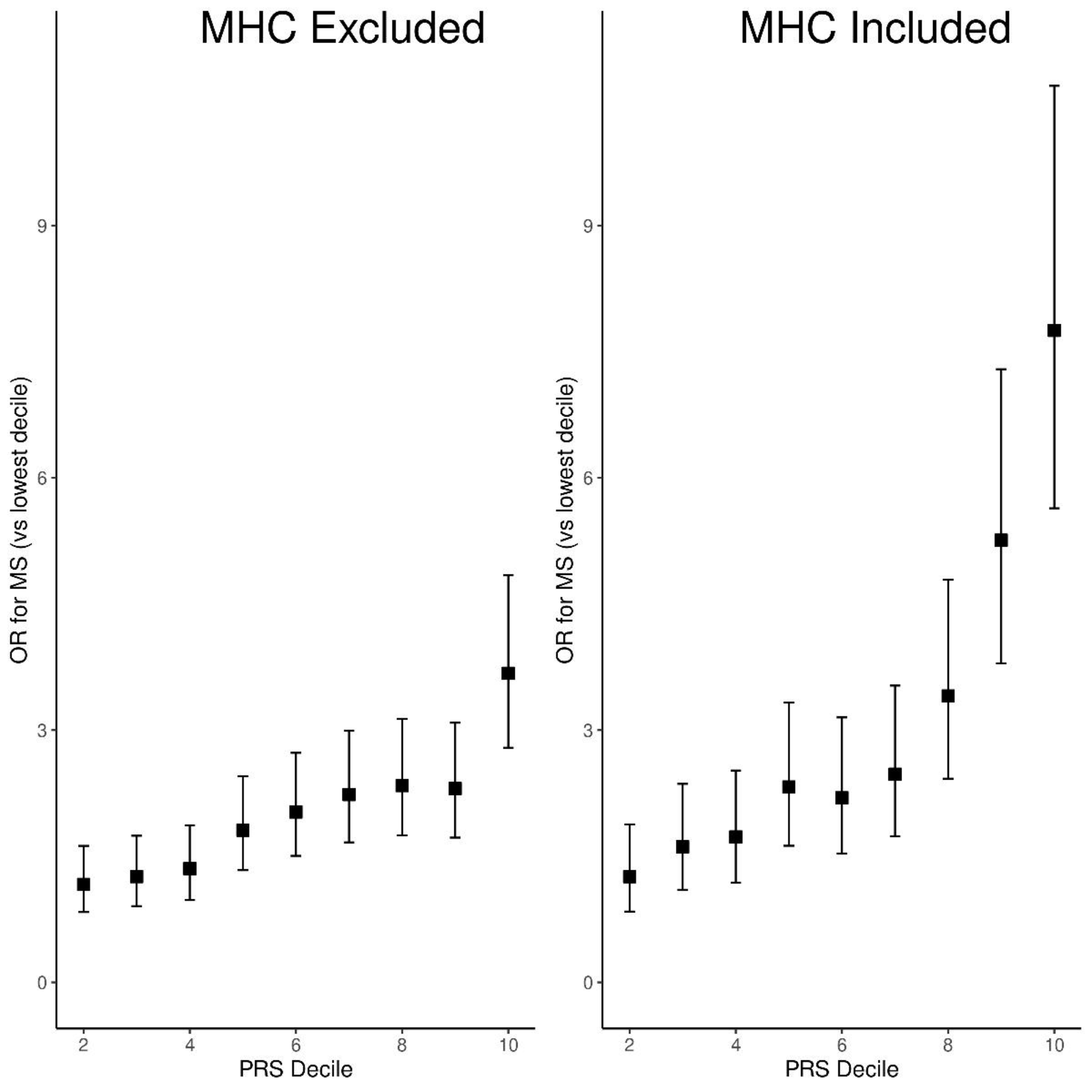
Figure 2: A: Nagelkerke's pseudo- R^2 metric for each of the individual PRS used. The R^2 was calculated by comparing the model fit (age, sex, Townsend deprivation index, the first 4 genetic PCs, and PRS) vs the null model (age, sex, Townsend deprivation index, and the first 4 genetic PCs). A variety 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. PRS are shown both including and excluding the MHC region. 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.

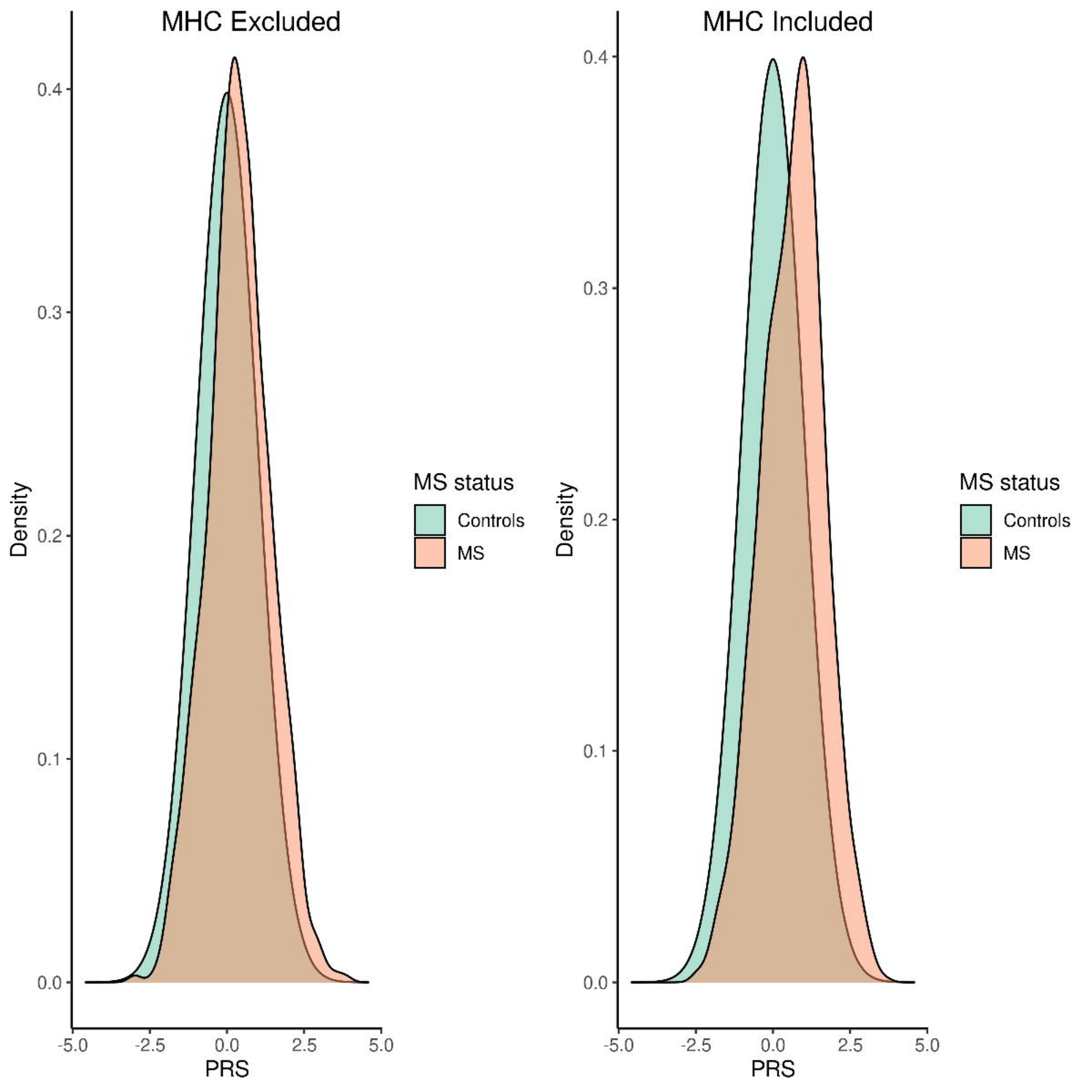
Figure 3: A: Calibration plot showing absolute MS disease probabilities within each PRS decile (of the non-MHC PRS). Other lines represent the mean fitted disease probabilities for models incorporating the MHC PRS, the non-MHC PRS, and null covariates only (age, sex, deprivation, genetic PCs). B: Receiver Operating Characteristic (ROC) curves demonstrating the discriminative performance (i.e. ability to distinguish MS cases from controls) of the null model, MHC PRS, and non-MHC PRS. C&D: scatter plots showing no relationship between PRS (MHC and non-MHC respectively) and normalised age at MS report.

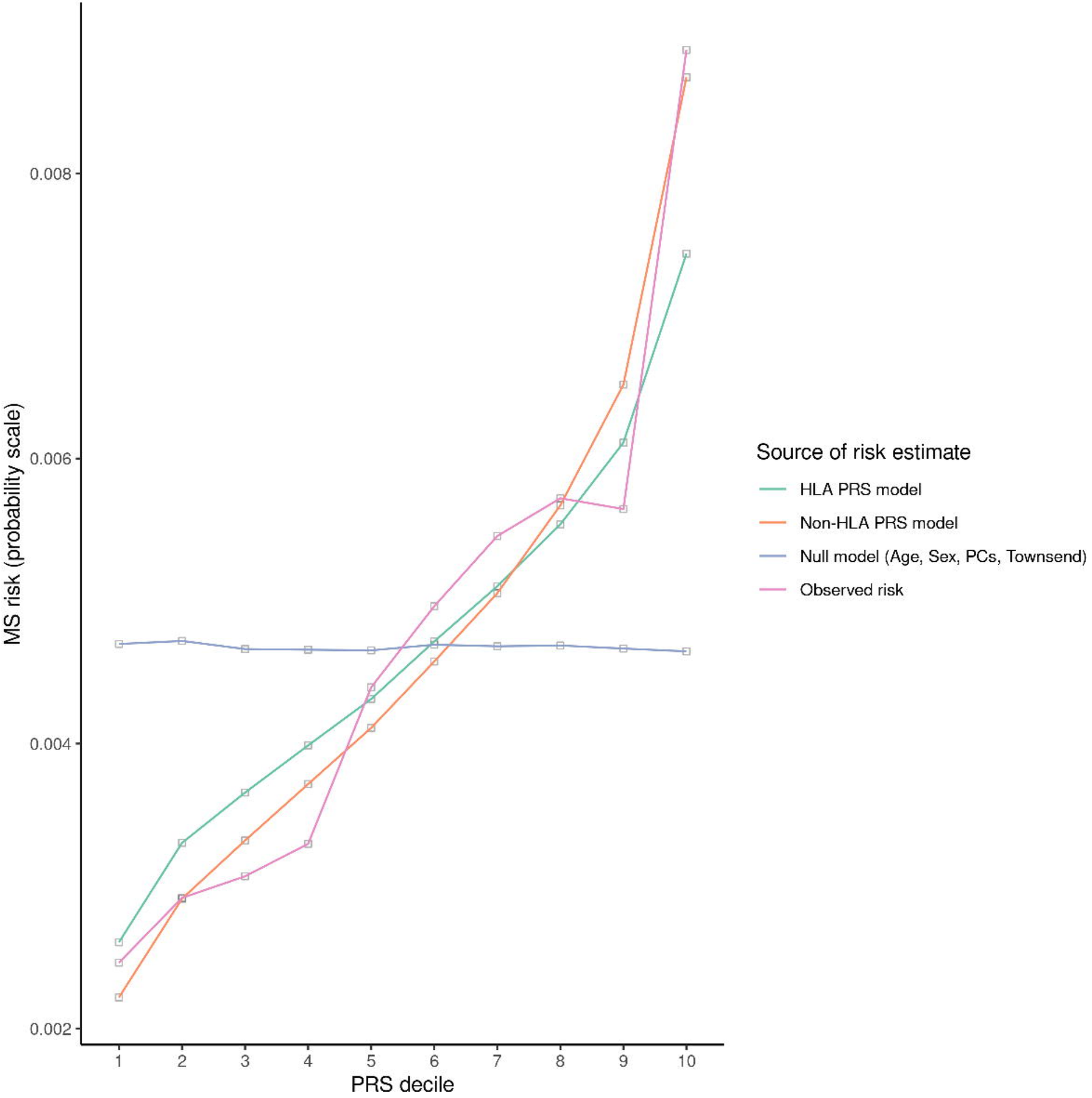
Figure 4: A: Forest plot demonstrating Attributable Proportion due to interaction (AP) and 95% CIs for interactions between environmental exposures and genetic risk factors for MS. If there is no interaction, the AP is 0. AP > 1 indicates positive interaction (combined effects exceed the sum of the individual effects), and vice-versa. 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 participants in the top and bottom PRS deciles, i.e. the highest and lowest 10% of polygenic risk scores. The outcome in each case is MS status, and the exposures of interest are childhood body size, age at menarche, smoking prior to age 20, and carriage of the HLA DRB1*15:01 allele. 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).

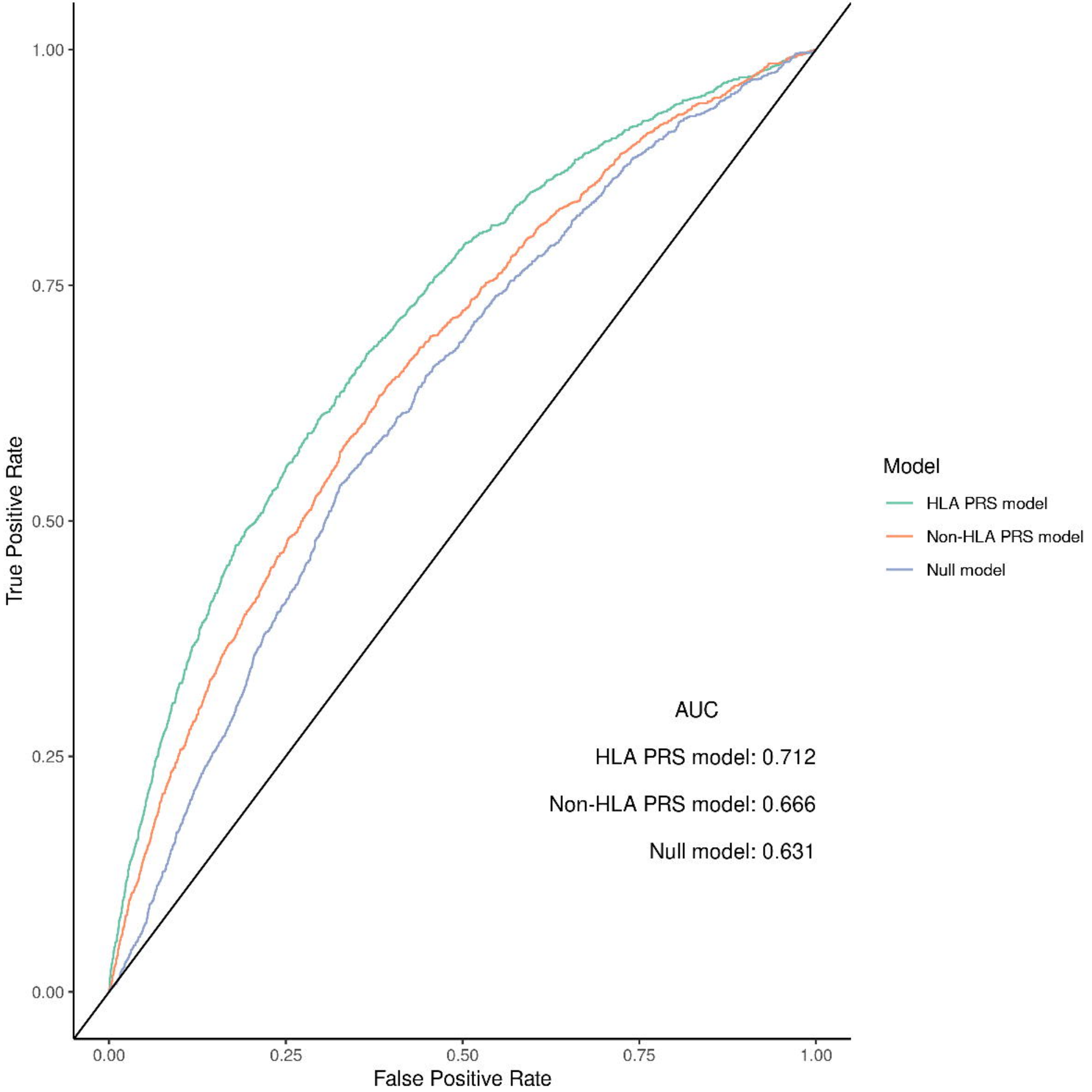

Exposed to maternal smoking (vs not)
Breastfed as a baby (vs not)
Born in September (vs April)
Born in October (vs April)
Born in November (vs April)
Born in May (vs April)
Born in March (vs April)
Born in June (vs April)
Born in July (vs April)
Born in January (vs April)
Born in February (vs April)
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)
Risk factor
-1
0
1
Log odds Ratio for MS (+/- 95% CI)

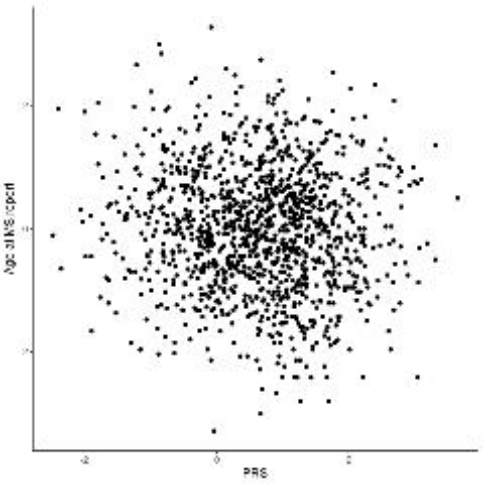

MHC excluded
MHC included
Nagelkerke Pseudo − R2
0.00
0.01
0.02
0.03
5e-08
0.05
0.1
0.2
0.4
0.6
0.8
1
P value threshold
Clumping R2 parameter
0.2
0.4
0.6
0.8

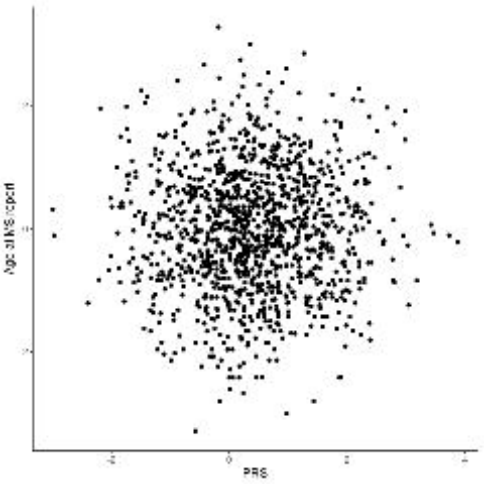

MHC Excluded
MHC Included
OR for MS (vs lowest decile)
0
3
6
9
2
4
6
8
10
PRS Decile

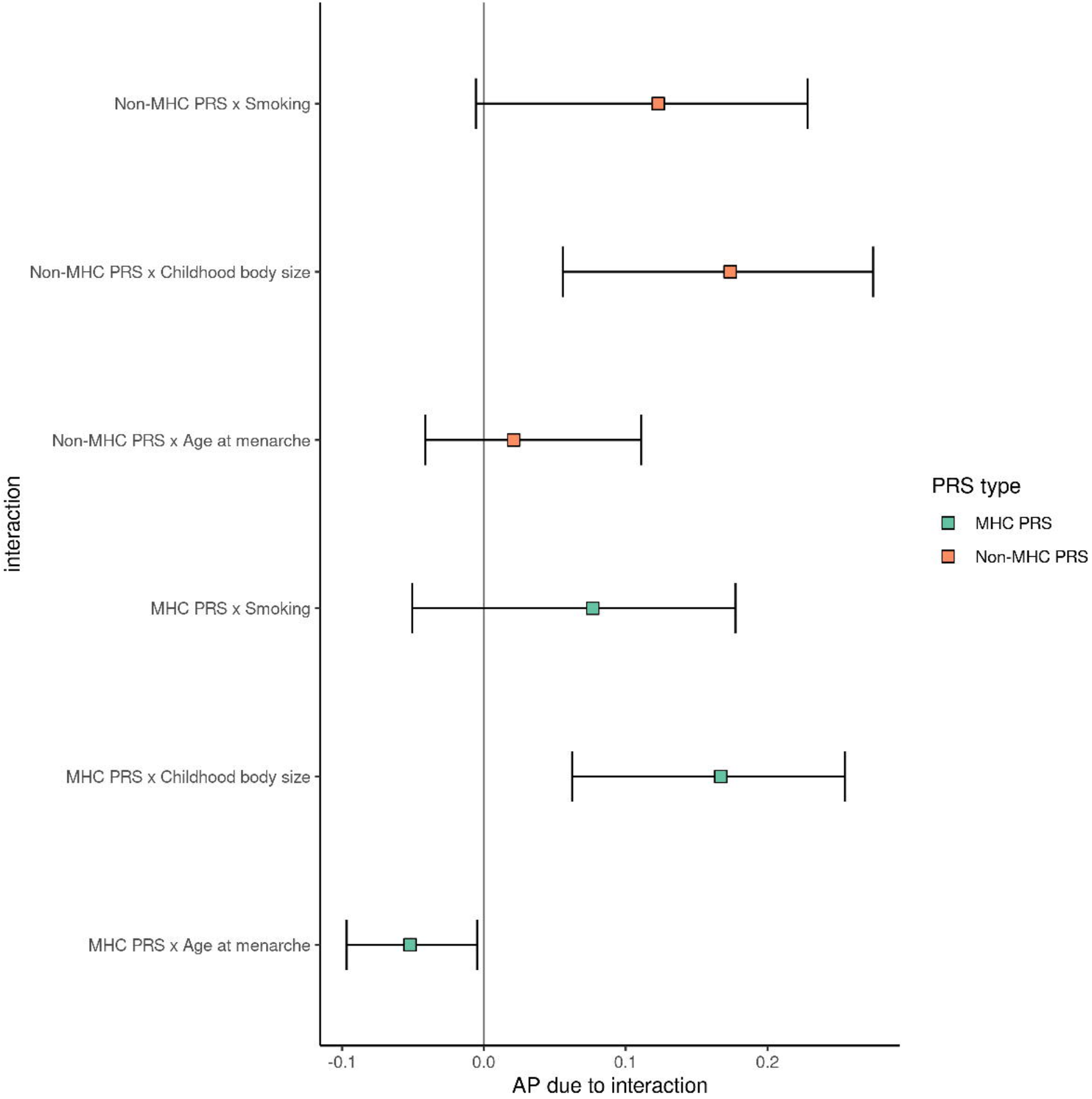

MHC Excluded
MHC Included
Density
PRS
0.0
0.1
0.2
0.3
0.4
-5.0
-2.5
0.0
2.5
5.0
MS status
Controls
MS

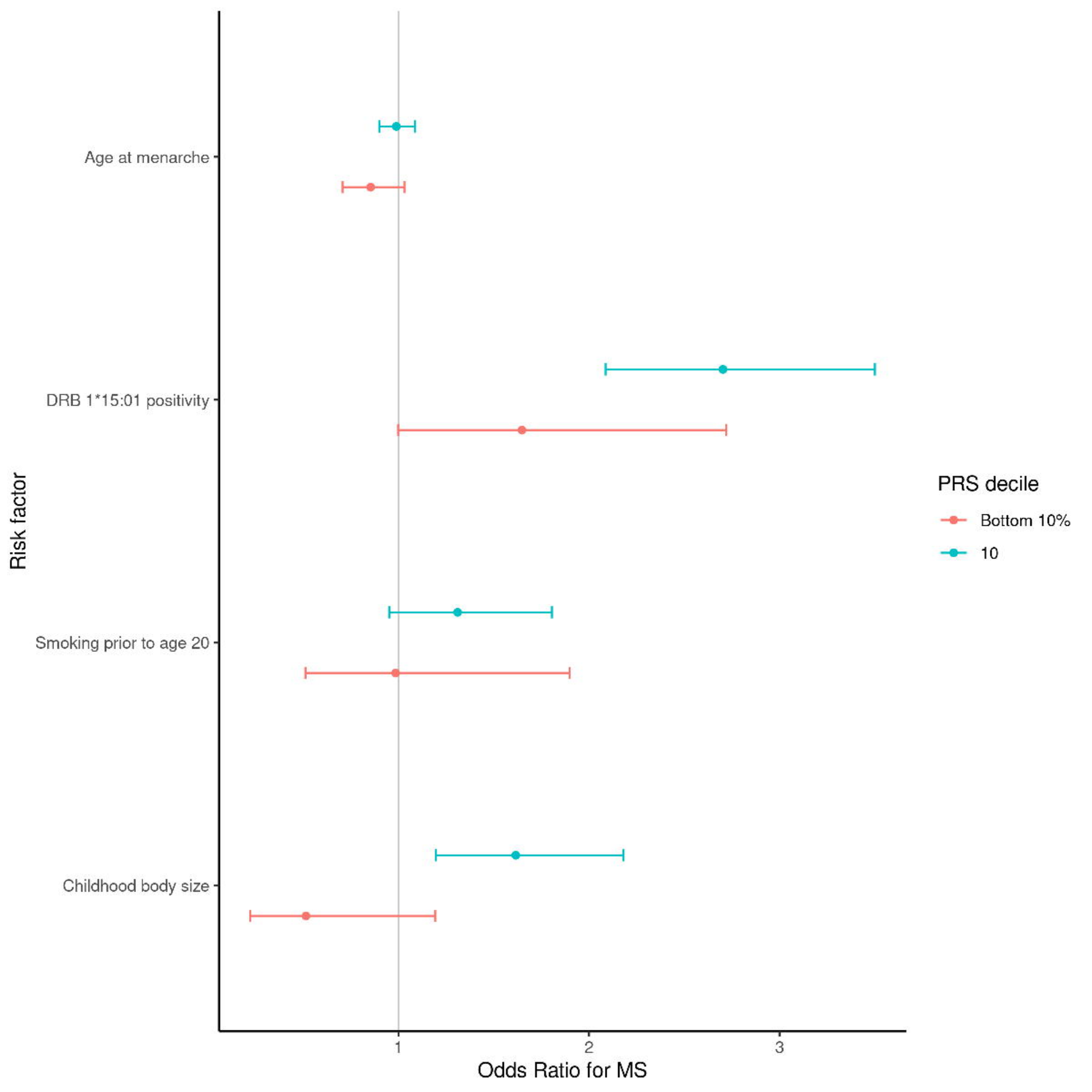

MS risk (probability scale)
0.008
0.006
0.004
0.002
1
2
3
4
5
6
7
8
9
10
PRS decile
Source of risk estimate
HLA PRS model
Non-HLA PRS model
Null model (Age, Sex, PCs, Townsend)
Observed risk


1.00
0.75
0.50
0.25
0.00
True Positive Rate
0.00
0.25
0.50
0.75
1.00
False Positive Rate
Model
HLA PRS model
Non-HLA PRS model
Null model
AUC
HLA PRS model: 0.712
Non-HLA PRS model: 0.666
Null model: 0.631


PRS


PRS


Non-MHC PRS x Smoking
Non-MHC PRS x Childhood body size
Non-MHC PRS x Age at menarche
MHC PRS x Smoking
MHC PRS x Childhood body size
MHC PRS x Age at menarche
interaction
-0.1
0.0
0.1
0.2
AP due to interaction
PRS type
MHC PRS
Non-MHC PRS


Age at menarche
DRB 1*15:01 positivity
Smoking prior to age 20
Childhood body size
Risk factor
1
2
3
Odds Ratio for MS
PRS decile
Bottom 10%
10