International Journal of Epidemiology, 2018, 1–10
doi: 10.1093/ije/dyy060
Original article

Original article

Exploring the association of genetic factors with participation in the Avon Longitudinal Study of Parents and Children

Amy E Taylor,[1,2,3,4] Hannah J Jones,[1,3,4] Hannah Sallis,[1,2,3] Jack Euesden,[1,3] Evie Stergiakouli,[1,3] Neil M Davies,[1,3] Stanley Zammit,[3,4,5] Debbie A Lawlor,[1,3,4] Marcus R Munafò,[1,2,4] George Davey Smith[1,3,4†] and Kate Tilling[1,3,4*†]

[1]MRC Integrative Epidemiology Unit at the University of Bristol, Bristol, United Kingdom, [2]UK Centre for Tobacco and Alcohol Studies, School of Experimental Psychology, University of Bristol, Bristol, United Kingdom, [3]Department of Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, United Kingdom, [4]National Institute for Health Research Bristol Biomedical Research Centre at the University Hospitals Bristol NHS Foundation Trust and the University of Bristol, Bristol, United Kingdom and [5]Institute of Psychological Medicine and Clinical Neurosciences, MRC Centre for Neuropsychiatric Genetics and Genomics, Cardiff University, Cardiff, United Kingdom

*Corresponding author. MRC IEU, University of Bristol, Oakfield House, Oakfield Grove, Bristol BS8 2BN, UK.
E-mail: kate.tilling@bristol.ac.uk
[†]These authors contributed equally to this work.

Editorial decision 18 March 2018; Accepted 4 April 2018

Abstract

Background: It is often assumed that selection (including participation and dropout) does not represent an important source of bias in genetic studies. However, there is little evidence to date on the effect of genetic factors on participation.

Methods: Using data on mothers ($N = 7486$) and children ($N = 7508$) from the Avon Longitudinal Study of Parents and Children, we: (i) examined the association of polygenic risk scores for a range of sociodemographic and lifestyle characteristics and health conditions related to continued participation; (ii) investigated whether associations of polygenic scores with body mass index (BMI; derived from self-reported weight and height) and self-reported smoking differed in the largest sample with genetic data and a subsample who participated in a recent follow-up; and (iii) determined the proportion of variation in participation explained by common genetic variants, using genome-wide data.

Results: We found evidence that polygenic scores for higher education, agreeableness and openness were associated with higher participation; and polygenic scores for smoking initiation, higher BMI, neuroticism, schizophrenia, attention-deficit hyperactivity disorder (ADHD) and depression were associated with lower participation. Associations between the polygenic score for education and self-reported smoking differed between the largest sample with genetic data [odds ratio (OR) for ever smoking per standard deviation (SD) increase in polygenic score: 0.85, 95% confidence interval (CI): 0.81, 0.89} and

subsample (OR: 0.96, 95% CI: 0.89, 1.03). In genome-wide analysis, single nucleotide polymorphism based heritability explained 18–32% of variability in participation.
Conclusions: Genetic association studies, including Mendelian randomization, can be biased by selection, including loss to follow-up. Genetic risk for dropout should be considered in all analyses of studies with selective participation.

Key words: ALSPAC, missing data, genetics, participation, selection bias

Key Messages

- Polygenic scores for a range of sociodemographic, health and lifestyle factors are related to continued participation after enrolment in the Avon Longitudinal Study of Parents and Children.
- There was evidence that associations between polygenic scores and measured phenotypes differed between the full sample with genetic data and a more selected subsample, indicating that genetic association studies can be biased by selection.
- Common genetic variation explained a moderate amount (18–32%) of variability in participation.
- Researchers should consider selective participation as a potential source of bias in genetic and non-genetic association studies.

Introduction

Missing data are a pervasive problem in cohort studies, with decreasing participation over the duration of the study, and concern about the extent to which this biases analyses.[1,2] Individual characteristics, including social and lifestyle characteristics, may influence both initial enrolment and continued participation.[3,4] Throughout this paper we use the word 'participation' to mean both initial enrolment in a study and continued participation (e.g. via questionnaire completion or attendance at research clinics) once involved. However, our analyses all relate to continued participation after enrolment.

Sample representativeness is critical for estimating prevalence of exposure or disease,[5] but may not be essential for estimating associations between exposures and outcomes.[5–7] The bias arising from selection into studies is often relatively small and may not always qualitatively affect interpretation of results.[1,8,9] Selection bias might be less problematic in genetic epidemiology because individuals are generally unaware of their genotype (so will not self-select into a study on the basis of this) and genetic variants that influence a given trait should not be associated with confounding factors which could also influence selection.[6,10] However, when both exposure and outcome relate to participation in a study, this can induce spurious associations between them, or between genetic variants that influence them, in participants.[11,12] For example, the association between higher genetic risk for schizophrenia and reduced participation in the Avon Longitudinal Study of Parents and Children (ALSPAC)[13] indicates that selection bias may be a problem in both genetic and non-genetic analyses of schizophrenia.

To estimate the impact of selective participation for a given analysis, we need to know which factors cause participation. Here, we extend previous work relating participation and polygenic risk for schizophrenia and autism in ALSPAC[13,14] by: (i) investigating polygenic scores for other factors which could influence participation in the ALSPAC mothers and children; (ii) investigating the potential impact of selection bias by comparing associations between genetic factors and measured phenotypes in the largest sample with genetic data and a more selected subsample; and (iii) conducting genome-wide association studies of participation measures.

Methods

Study population

ALSPAC is a longitudinal birth cohort that recruited 14 541 pregnant women resident in Avon, UK, with expected dates of delivery between 1 April 1991 and 3 December 1992. Of these initial pregnancies, there were a total of 14 676 fetuses, resulting in 14 062 live births and 13 988 children who were alive at 1 year of age. The children and their mothers have been followed up through postal questionnaires and at clinics.[3,15] We included only children who had been enrolled in the study during the first phase of data collection and survived to age 1 year (resulting in the exclusion of five children

and 43 mothers from the analysis sample). Please note that the study website contains details of all the data that are available through a fully searchable data dictionary: [http://www.bris.ac.uk/alspac/researchers/data-access/data-dictionary]. Ethical approval for the study was obtained from the ALSPAC Ethics and Law Committee and the local research ethics committees.

Participation

Participation was defined by responding to a questionnaire or attending a clinic for which the whole cohort was eligible to participate (i.e. we excluded clinics and questionnaires targeted at a subset of the cohort). The ALSPAC mothers have answered questionnaires about themselves (mother questionnaires) and about their children (child-based questionnaires). The ALSPAC children have answered questionnaires about themselves (child-completed questionnaires). A full list of the questionnaires and clinics included is provided in Supplementary Table 1, available as Supplementary data at *IJE* online. From these, we calculated the following continuous phenotypes by summing the number of questionnaires/clinics completed: total participation [all questionnaires and clinics for both mother and child (including child-based and child-completed)]; total questionnaire (all questionnaires for mothers and children); mother questionnaire (mother questionnaires); child questionnaire (child-completed questionnaires); and child clinic (child clinics attended). We created two binary variables for the mothers and children indicating: (i) participation in the most recent clinic; and (ii) completion of the most recent questionnaire. For both mothers and the offspring, we generated variables from data collected at clinics 17–18 years after the child's birth and from questionnaires 19–20 years after birth.

Genetic data

ALSPAC children were genotyped using the Illumina HumanHap550 quad chip genotyping platforms. ALSPAC mothers were genotyped using the Illumina Human660W-quad array at the Centre National de Genotypage (CNG), and genotypes were called with Illumina GenomeStudio. Imputation was performed using Impute V2.2.2 against the 1000 genomes phase 1 version 3 reference panel. Quality control procedures removed related individuals and individuals of non-European genetic ancestry (see Supplementary materials for full details, available as Supplementary data at *IJE* online).

Polygenic scores

We calculated polygenic scores for a number of traits that could be related to participation and for which genome-wide summary statistics were publicly available: body mass index,[16] height,[17] smoking initiation,[18] depression,[19] attention-deficit hyperactivity disorder (ADHD),[20] bipolar disorder,[21] autism,[21] schizophrenia,[22] years of education,[23] sleep duration,[24] chronotype (morningness),[24] age at menarche,[25] personality traits (openness, agreeableness, conscientiousness, extraversion and neuroticism)[26] and Alzheimer's disease.[27] For the purposes of this paper, we use the term 'trait' to describe the phenotype each genome-wide association study (GWAS) was conducted on but acknowledge that, for binary phenotypes, we are looking at genetic liability for that phenotype. Full details of sources for each of these scores are shown in Supplementary Table 2, available as Supplementary data at *IJE* online. The ALSPAC cohort was not included in the GWAS that generated the summary statistics for these traits, except for education and age at menarche. For education, we used summary statistics excluding ALSPAC and 23andme, which were obtained directly from the study authors. For age at menarche, the ALSPAC sample made up 7% of the GWAS discovery sample.[25] To minimize potential bias from sample overlap, we used an unweighted polygenic score for age at menarche.[28] All other scores were weighted according to the association magnitude of each single nucleotide polymorphism (SNP) in the original GWAS.

Statistical analysis

All analyses were performed separately in mothers and children and were adjusted for sex (in the children) and the first 10 genetic principal components.

Polygenic scores

Polygenic scores were derived using the PRSice software [http://prsice.info/][29] for each trait within the ALSPAC genome-wide data using the following *P*-value thresholds: 0.0005, 0.005, 0.05, 0.1, 0.5 (see Supplementary Methods, available as Supplementary data at *IJE* online). In addition, we generated scores in PRSice by inputting only the independent genome-wide significant SNPs reported by the discovery samples (Supplementary Table 3, available as Supplementary data at *IJE* online). We assessed associations of standardized polygenic scores with participation phenotypes using linear and logistic regression in Stata (version 14.1).[30] We used robust standard errors to account for the non-normal distribution of the continuous participation variables. For age at menarche, analyses were conducted in females only.

Genome-wide association analysis

Analyses were conducted separately for mothers and children. We used SNPTEST[31] to test associations between

Table 1. Summary of participation phenotypes

	Range	Mother ($N = 7486$) Median (IQR)	Child ($N = 7508$) Median (IQR)
Total participation	0–77	59 (31,71)	62 (39,72)
Total questionnaire	0–67	53 (29,63)	55 (35,63)
Mother questionnaire	0–19	16 (10,18)	–
Child questionnaire	0–24	–	17 (8,22)
Child clinic	0–9	–	7 (3,9)
		N (%)	*N* (%)
Mother attended most recent clinic		3215 (43.0)	–
Mother completed most recent questionnaire		3052 (40.8)	–
Child attended most recent clinic		–	3538 (47.1)
Child completed most recent questionnaire		–	2957 (39.4)

IQR, interquartile range.

dosage scores for each genetic variant and missingness phenotypes using univariate regression models and assuming an additive genetic model. Continuous phenotypes were initially tested in linear models, and then dichotomized at arbitrary midpoints (Supplementary Table 4, available as Supplementary data at *IJE* online) and re-tested in logistic models to ensure results were robust to any assumption on the distribution of residuals. Genome-wide results were filtered to remove SNPs with a minor allele frequency of <0.01 and imputation quality (info) score of <0.8. Genome-wide significance was considered to be $P < 5 \times 10^{-8}$.[32]

Heritability

SNP-based heritability estimates h^2_{SNP} were calculated for each participation phenotype using the genetic restricted maximum likelihood (GREML) method implemented within the GCTA software.[33]

Investigating the impact of selection bias in ALSPAC

We used linear and logistic regression to calculate associations between polygenic scores for BMI, smoking, education and schizophrenia (constructed at a*P*-value threshold of 0.05) and body mass index and smoking status (ever vs never smoking) which were self-reported by the ALSPAC mothers in questionnaires administered during pregnancy. These analyses were conducted first in the largest sample with genome-wide data and then in the sample attending the most recent clinic.

Results

Of the 13 793 mothers with 13 988 children alive at 1 year, 11 560 mothers and 10 780 children had provided DNA samples. After removal of non-Europeans, related individuals and samples which did not pass quality control, 7486 mothers and 7508 children were eligible for analysis (Table 1, Supplementary Figures 1 and 2, available as Supplementary data at *IJE* online). Individuals included in the analysis had higher participation levels than the enrolled cohort (Supplementary Table 5, available as Supplementary data at *IJE* online). Continuous participation phenotypes were highly correlated (Pearson's correlation coefficients ranged between 0.71 and 0.99) (Supplementary Table 6, available as Supplementary data at *IJE* online).

Associations of polygenic scores with participation phenotypes

Only the results for total participation and last questionnaire completion are presented, with results for all other participation measures in Supplementary material, available as Supplementary data at *IJE* online.

In ALSPAC mothers, we found strong evidence for positive associations between polygenic scores for years of education and participation. This was observed consistently across all participation phenotypes (Figures 1 and 2, and Supplementary Figures 3–5, available as Supplementary data at *IJE* online). Higher values of polygenic scores for height and agreeableness were also associated with higher participation across most participation phenotypes. There was also some evidence that higher polygenic scores for openness were associated with the mother completing more questionnaires about herself. In contrast, polygenic scores for BMI, schizophrenia, ADHD, smoking initiation and depression were negatively associated with participation. Polygenic scores for neuroticism were associated with lower participation by the mothers.

Associations between polygenic scores and participation were similar for ALSPAC children (Figures 3 and 4, and

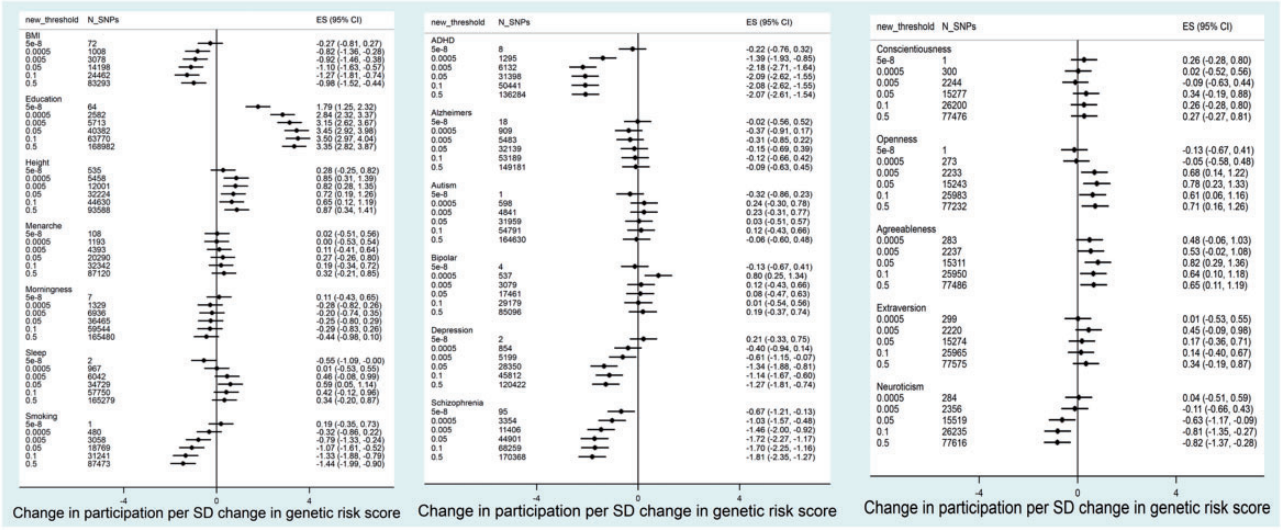

Figure 1. Association between polygenic scores in ALSPAC mothers and total participation score ($N = 7468$).

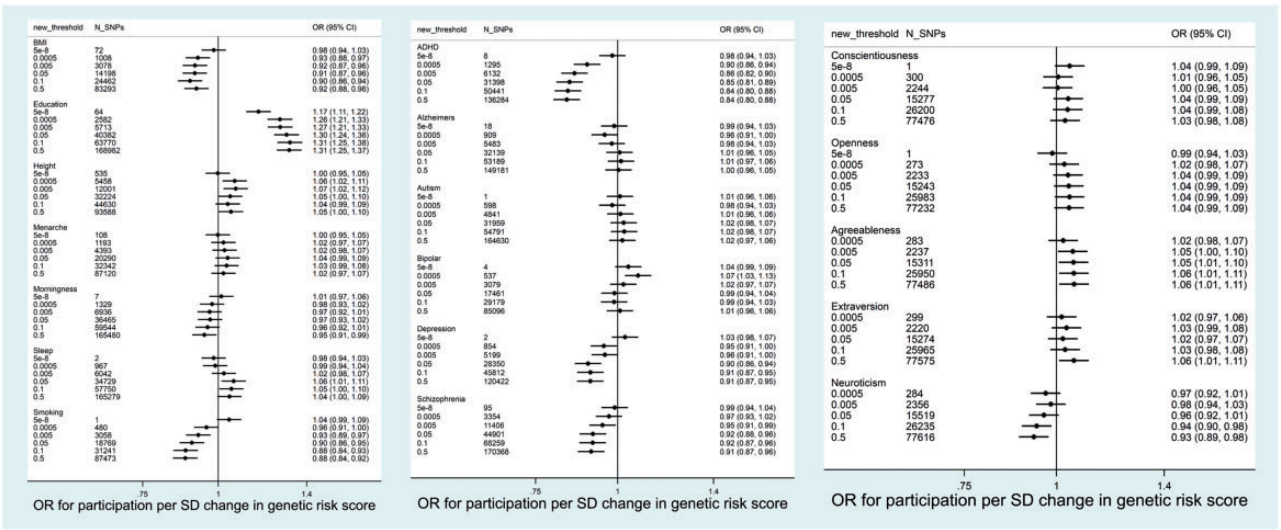

Figure 2. Association between polygenic scores in ALSPAC mothers and completion of most recent questionnaire ($N = 7468$).

Supplementary Figures 6–9, available as Supplementary data at *IJE* online). Polygenic scores for education and agreeableness were positively associated with participation. Polygenic scores for smoking initiation, schizophrenia, ADHD and depression were negatively associated with participation. In contrast to the ALSPAC mothers, there was little evidence for associations between polygenic scores for neuroticism, height or openness and participation.

We found no consistent evidence that polygenic scores for morningness (chronotype), sleep, bipolar disorder, autism, conscientiousness, extraversion, age at menarche or Alzheimer's disease were associated with participation.

Correlations between polygenic scores

The degree of correlation between polygenic scores for different traits at $P < 0.0005$ and $P < 0.5$ is shown in Supplementary Tables 7–10, available as Supplementary data at *IJE* online. Correlations tended to be stronger for scores derived using the higher *P*-value thresholds.

Investigating the impact of selection bias in ALSPAC

Figure 5 shows associations (in the largest sample with genome-wide data and in a subsample who attended the

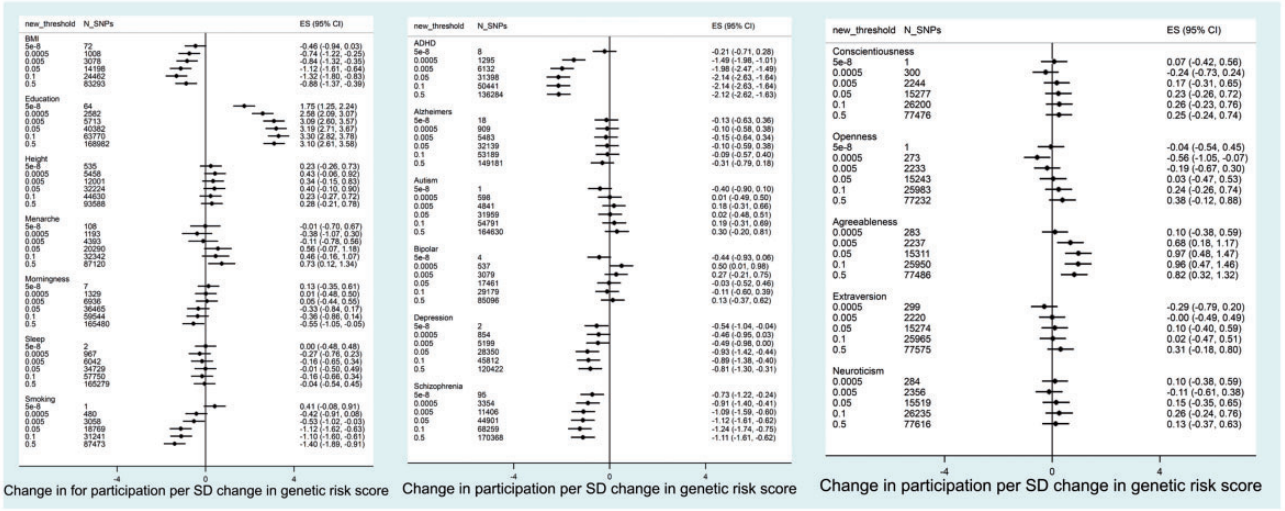

Figure 3. Association between polygenic scores in ALSPAC children and total participation score ($N = 7508$). Age at menarche analysis only in females.

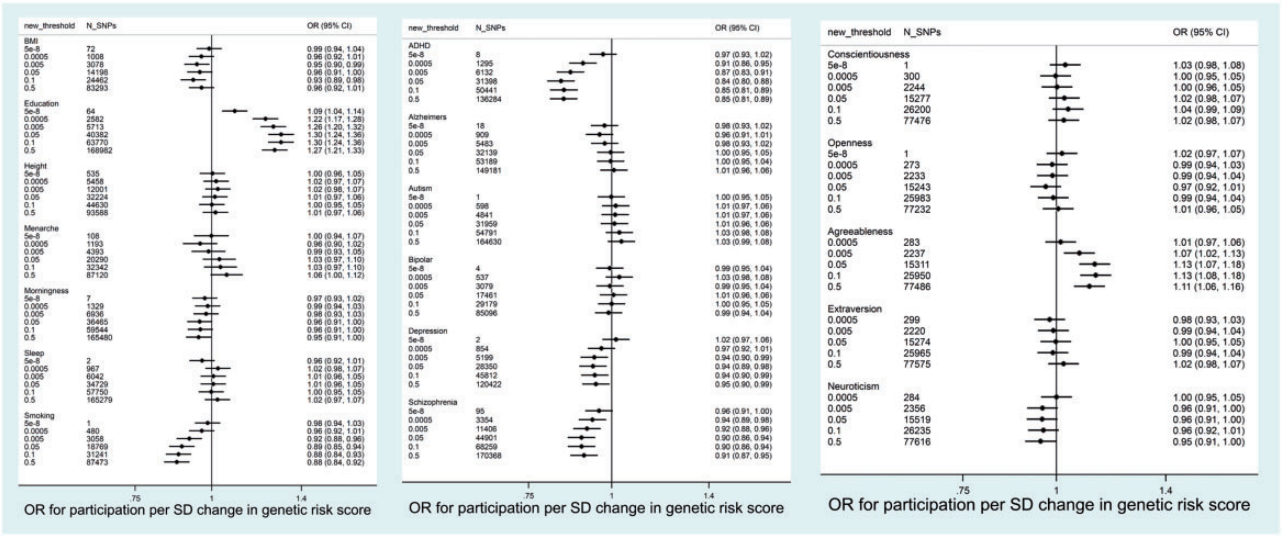

Figure 4. Association between polygenic scores in ALSPAC children and completion of most recent questionnaire ($N = 7508$). Age at menarche analysis only in females.

most recent clinic) between polygenic scores (constructed at the $P < 0.05$ threshold) for BMI, smoking, education and schizophrenia and self-reported BMI and smoking. Associations between each polygenic score and smoking or BMI were in the same direction in both the full sample and the subsample, and in many cases of similar magnitude. However, associations between the polygenic score for education and being an ever smoker were substantially attenuated in the subsample [odds ratio (OR): 0.96 per standard deviation (SD) in polygenic score for smoking, 95% confidence interval (CI): 0.89, 1.03, compared with the full genetic sample (OR: 0.85, 95% CI: 0.81, 0.89)] (Figure 5A). The association between the education polygenic score and BMI was also attenuated in the subsample compared with the full sample (Figure 5B). In contrast, the association between the smoking polygenic score and BMI appeared stronger in the subsample compared with the full genetic sample, although the confidence intervals overlapped.

Genome-wide association studies

Only one locus reached genome-wide significance with participation in the ALSPAC mothers. In the mothers, variants located in an intergenic region on chromosome 7: 51995163–52042976 were associated with total participation, total questionnaire and mother questionnaire (Figure 6, Supplementary Figures 10–11 and Supplementary

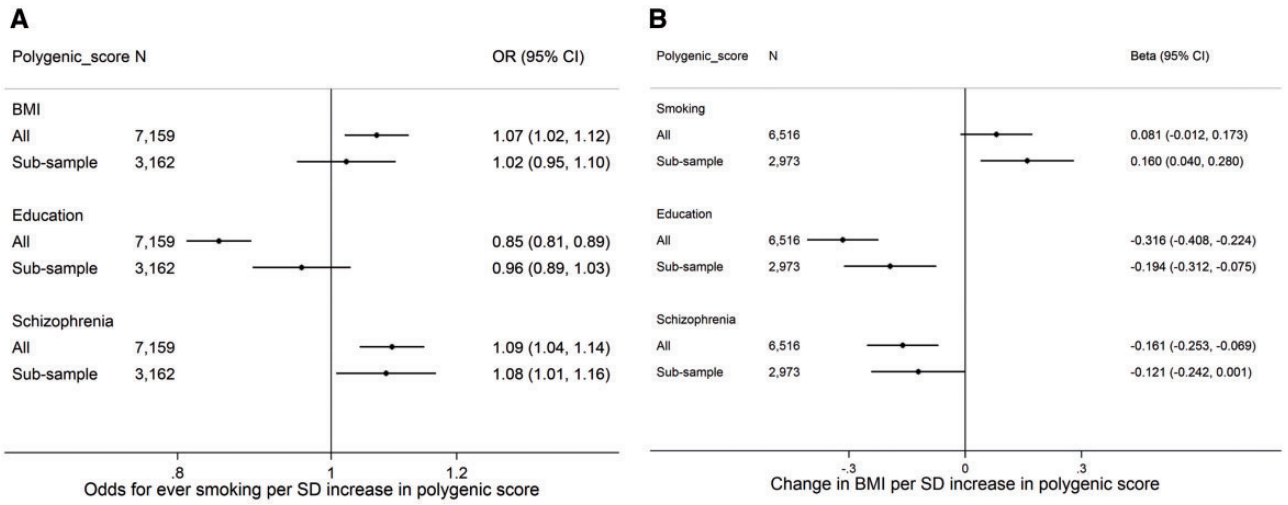

Figure 5. Association between genetic risk scores for BMI, smoking, education and schizophrenia, and self-reported smoking and BMI, conditioned on attendance at the most recent clinic. Analyses adjusted for first 10 genetic principal components.

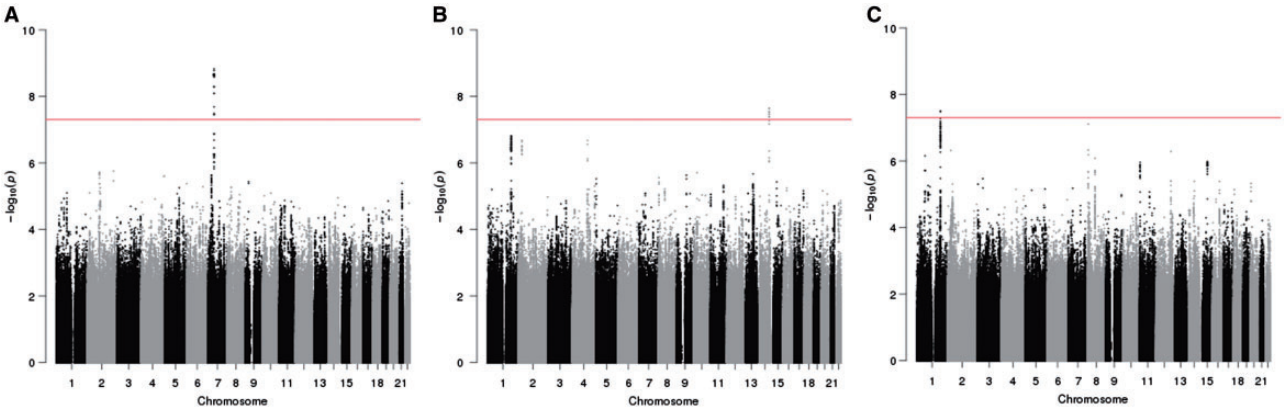

Figure 6. Manhattan plots for genome-wide analyses of total participation in the mothers and children and clinic participation in the children. A. Total participation score in the mothers, B. Total participation score in the children, C. Number of clinics attended by the child. Line represents $P = 5 \times 10^{-8}$.

Tables 11–13, available as Supplementary data at *IJE* online). Genome-wide hits were all in strong linkage disequilibrium ($R^2 > 0.8$), indicating that this represents a single genetic signal. The SNP with the smallest *P*-value was rs10626545 for total ($P = 1.50 \times 10^{-9}$) and total questionnaire ($P = 8.55e^{-10}$), and rs406001 for mother questionnaire ($P = 8.27 \times 10^{-9}$). SNPs in this region reached genome-wide significance or close to genome-wide significance ($P < 7 \times 10^{-7}$) with dichotomized total participation, total questionnaire and mother questionnaire (data not shown). However, the minor allele frequency of these variants was relatively low (0.012) and beta-coefficients large (beta for total participation for top SNP= 10.9), suggesting that this association is driven by a few individuals.

In the children, two loci reached genome-wide significance (Figure 6, Supplementary Figures 12–13 and Supplementary Tables 14–16, available as Supplementary data at *IJE* online). SNPs in the bradykinin receptor B1 gene (*BDKRB1)* (chromosome 14: 96721850–96729885) were associated with total participation, total questionnaire and child questionnaire. The SNP with the smallest *P*-value was rs28631073 for all three participation measures (*P* between 1.29×10^{-8} and 2.27×10^{-8}) and the beta with total participation was −3.20. Two SNPs in an intergenic region on chromosome 1 reached genome-wide significance with child clinic participation: rs1336852 (1: 191752825, beta: −0.59, $P = 3.15 \times 10^{-8}$) and rs74626786 (1: 191759598, beta: −0.59, $P = 3.32 \times 10^{-8}$). Plots showing linkage disequilibrium and nearest genes for each of the genome wide significant loci (created using LocusZoom[34]) are shown in Supplementary material (Figures 14–20, available as Supplementary data at *IJE* online).

SNP-based heritability

Estimates of heritability of participation phenotypes from SNPs included in the genome-wide analyses ranged 20–27% for the mothers and 18–32% for the children

(*P*-values all <0.001) (Supplementary Table 17, available as Supplementary data at *IJE* online).

Discussion

Continued participation in the ALSPAC cohort is related to polygenic scores for a number of lifestyle factors, personal characteristics and health conditions, including level of education, BMI, height, smoking, agreeableness, openness, schizophrenia, ADHD and depression. We did not find robust evidence in genome-wide analyses that specific single genetic variants influence degree of participation in ALSPAC, though there was evidence of common genetic variants explaining a modest proportion of the variation in participation (up to 30%).

Our findings show that genetic variants which are related to specific phenotypes are also related to participation. Using a Mendelian randomization framework, this could imply that these phenotypes cause continued participation. For example, the polygenic risk score for education was the score most robustly associated with participation—implying that higher education causes greater continued participation in ALSPAC. This interpretation requires that the key assumptions of Mendelian randomization are met,[35] namely that: (i) the polygenic score is robustly associated with the trait of interest; (ii) there are no confounders of the polygenic score-participation association; and (iii) the genetic risk score only affects participation through the trait of interest. The third of these assumptions is more likely to be met as the threshold for polygenic score construction gets closer to genome-wide significance.

Polygenic scores created using higher *P*-value thresholds could explain more of the variance in that trait than genome-wide significant scores,[36] but are likely to be less specific for the trait of interest and more likely to be pleiotropic, influencing more than one trait. This is shown by the stronger correlations between risk scores for different traits created at high *P*-value thresholds than those created using low *P*-value thresholds. We found traits for which genome-wide scores were not associated with participation, but scores at higher *P*-value thresholds were, for example depression. This could be explained by low power in the original GWAS, meaning that truly associated SNPs are less likely to be included in a score constructed using a low significance threshold,[37] or that effects on participation are acting through a trait that is only distally related to the GWAS trait used in score construction. As the *P*-value threshold increases, this also introduces more noise into the polygenic scores and may explain why some scores at the $P = 0.5$ threshold are less strongly associated with participation than the scores created at lower thresholds.

We also showed that it is possible to introduce bias into genetic analyses even when sample sizes are relatively modest. Therefore, we cannot assume that genetic-association studies, including GWAS, candidate gene studies and Mendelian randomization, are not biased by incomplete participation. We recommend that researchers consider how likely non-participation is as a potential source of bias when running genetic association studies and acknowledge this when reporting findings. The same implications hold for non-genetic studies—e.g. a study of the association between education levels and BMI in a selected subsample is likely to be biased by selection, since our genetic results show that both exposure and outcome cause participation.

For both genetic and non-genetic studies, there are potential methods to correct for this bias. For example, where there is some information about participants who have dropped out, it may be possible to apply inverse probability weighting.[38] Where such data are not available, other approaches could be triangulated to examine likelihood of bias. Negative control exposures and/or outcomes can be used to see if associations between genetic variants and outcomes exist that are not biologically plausible and should only arise through selection bias.[39] Similarly, where there is a well characterized association (replicated in a number of studies) of known magnitude between a genetic variant and an outcome, this can be used as a positive control. Finally, novel associations should be replicated in populations which have not undergone the same degree of selection.

We found three loci associated with participation at genome-wide significance level. SNPs in the genomewide locus in mothers (e.g. rs406001) were identified in a previous GWAS of post-traumatic stress disorder (PTSD), but not replicated in the original GWAS.[40] Furthermore, this locus was only nominally associated with PTSD in a much larger GWAS.[41] This, coupled with the low minor allele frequency of SNPs in the genome-wide significant locus in our GWAS, suggests that this may be a chance finding, rather than an effect of PTSD on participation. The signal on chromosome 14 is located in the bradykinin receptor B1 gene (*BDKRB1)*. Bradykinin is a peptide hormone which is a pro-inflammatory mediator and is involved in vascular permeability and mitogenesis.[42] To our knowledge, variants in this gene and the genome-wide significant SNPs on chromosome 1 have not been identified in previous GWAS of any phenotype.[43,44] We have not attempted to replicate the genome-wide hits in independent samples, as we cannot assume that different studies would have the same influences on participation.

There are a number of limitations to this analysis. First, our analysis sample was restricted to just over half of the

enrolled sample, due to availability of DNA samples for GWAS and exclusion criteria (non-Europeans and related individuals). Individuals in the analysis sample had higher participation rates than the full sample, meaning that associations between polygenic scores and participation are likely to be weaker than we would observe if we had full genetic data for the whole cohort. Second, our results may not be generalizable to studies with different selection criteria or specific cultural or contextual factors influencing participation. It is also possible that characteristics influencing participation will change over time and with age. We have shown here that genetic associations can be used to shed light on the selection mechanisms operating in a given study, but this will need repeating in studies in different populations or with different recruitment mechanisms. These are context-specific, rather than biological associations—although there is evidence that some associations (e.g. with education) may be fairly replicable.[45] Third, we have not attempted to disentangle the relative influence of maternal and offspring genetics on participation. It is likely that child participation is heavily influenced by maternal traits in childhood and this may continue into adolescence and adulthood. Finally, we have not explored all possible traits that might be associated with participation, since our analyses required access to GWAS summary statistics.

In conclusion, we demonstrate that polygenic scores related to a wide range of traits are associated with degree of participation in ALSPAC, and that this may introduce bias into genetic and non-genetic analyses. This highlights the importance of considering selection bias in all studies, and the need for the development of statistical methods to account for this issue.

Supplementary Data

Supplementary data are available at *IJE* online.

Funding

The UK Medical Research Council and Wellcome (Grant ref: 102215/2/13/2) and the University of Bristol provide core support for ALSPAC. Funding from British Heart Foundation, Cancer Research UK, Economic and Social Research Council, UK Medical Research Council and the National Institute for Health Research, under the auspices of the UK Clinical Research Collaboration, is gratefully acknowledged. This work was supported by the UK Medical Research Council (MC_UU_12013/1, MC_UU_12013/5, MC_UU_12013/6, MC_UU_12013/9, MR/M006727/1). This study was supported by the NIHR Biomedical Research Centre at University Hospitals Bristol NHS Foundation Trust and the University of Bristol. The views expressed in this publication are those of the author(s) and not necessarily those of the the MRC, Wellcome, the NHS, the National Institute for Health Research or the Department of Health.

Acknowledgements

We are extremely grateful to all the families who took part in this study, the midwives for their help in recruiting them, and the whole ALSPAC team, which includes interviewers, computer and laboratory technicians, clerical workers, research scientists, volunteers, managers, receptionists and nurses. ALSPAC children were genotyped using the Illumina HumanHap550 quad chip genotyping platforms by Sample Logistics and Genotyping Facilities at Wellcome Sanger Institute and LabCorp (Laboratory Corporation of America), using support from 23andMe.

Conflicts of interest: A.E.T. and M.R.M. are members of the UK Centre for Tobacco Control Studies, a UKCRC Public Health Research: Centre of Excellence. A.T. is in receipt of a GRAND award from Pfizer for research which is unrelated to the current submission. J.E. currently works for GlaxoSmithKline. D.A.L. has received support from National and International government and health charity research funders and also from Roche Diagnostics and Medtronic for research unrelated to that presented here.

References

1. Howe LD, Tilling K, Galobardes B, Lawlor DA. Loss to follow-up in cohort studies: bias in estimates of socioeconomic inequalities. *Epidemiology* 2013;**24**:1–9.
2. Powers J, Loxton D. The impact of attrition in an 11-year prospective longitudinal study of younger women. *Ann Epidemiol* 2010;**20**:318–21.
3. Boyd A, Golding J, Macleod J, Lawlor DA, Fraser A, Henderson J. Cohort Profile: The 'Children of the 90s'—the index offspring of the Avon Longitudinal Study of Parents and Children. *Int J Epidemiol* 2013;**42**:111–27.
4. Lamers F, Hoogendoorn AW, Smit JH *et al*. Sociodemographic and psychiatric determinants of attrition in the Netherlands Study of Depression and Anxiety (NESDA). *Compr Psychiatry* 2012;**53**:63–70.
5. Elwood JM. Commentary: On representativeness. *Int J Epidemiol* 2013;**42**:1014–15.
6. Ebrahim S, Davey Smith G. Commentary: Should we always deliberately be non-representative? *Int J Epidemiol* 2013;**42**:1022–26.
7. Rothman KJ, Gallacher JE, Hatch EE. Why representativeness should be avoided. *Int J Epidemiol* 2013;**42**:1012–14.
8. Pizzi C, De Stavola B, Merletti F *et al*. Sample selection and validity of exposure-disease association estimates in cohort studies. *J Epidemiol Community Health* 2011;**65**:407–11.
9. Pizzi C, De Stavola BL, Pearce N *et al*. Selection bias and patterns of confounding in cohort studies: the case of the NINFEA web-based birth cohort. *J Epidemiol Community Health* 2012;**66**:976–81.
10. Davey Smith G. The Wright stuff: genes in the interrogation of correlation and causation. *Eur J Pers* 2012;**26**:395–97.
11. Munafo MR, Tilling K, Taylor AE, Evans DM, Davey Smith G. Collider scope: when selection bias can substantially influence observed associations. *Int J Epidemiol* 2018;**47**:226–35.

12. Paternoster L, Tilling K, Davey Smith G. Genetic epidemiology and mendelian randomization for informing disease therapeutics: conceptual and methodological challenges. *PLoS Genet* 2017;**13**: e1006944.
13. Martin J, Tilling K, Hubbard L *et al.* Genetic risk for schizophrenia associated with non-participation over time in a population-based cohort study. *Am J Epidemiol* 2016;**183**:1149–58.
14. St Pourcain B, Robinson EB, Anttila V *et al.* ASD and schizophrenia show distinct developmental profiles in common genetic overlap with population-based social communication difficulties. *Mol Psychiatry* 2018;**23**:263–70.
15. Fraser A, Macdonald-Wallis C, Tilling K *et al.* Cohort Profile: The Avon Longitudinal Study of Parents and Children: ALSPAC mothers cohort. *Int J Epidemiol* 2013;**42**:97–110.
16. Locke AE, Kahali B, Berndt SI *et al.* Genetic studies of body mass index yield new insights for obesity biology. *Nature* 2015;**518**: 197–206.
17. Wood AR, Esko T, Yang J *et al.* Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat Genet* 2014;**46**:1173–86.
18. Furberg H, Kim Y, Dackor J *et al.* Genome-wide meta-analyses identify multiple loci associated with smoking behavior. *Nat Genet* 2010;**42**:441–47.
19. Okbay A, Baselmans BM, De Neve JE *et al.* Genetic variants associated with subjective well-being, depressive symptoms, and neuroticism identified through genome-wide analyses. *Nat Genet* 2016;**48**:624–33.
20. Demontis D, Walters RK, Martin J *et al.* Discovery of the first genome-wide significant risk loci for ADHD. *BioRxiv* 2017; 145581. doi: 10.1101/145581.
21. Cross-Disorder Group of the Psychiatric Genomics Consortium. Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. *Lancet* 2013;**381**: 1371–79.
22. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* 2014;**511**:421–27.
23. Okbay A, Beauchamp JP, Fontana MA *et al.* Genome-wide association study identifies 74 loci associated with educational attainment. *Nature* 2016;**533**:539–42.
24. Jones SE, Tyrrell J, Wood AR *et al.* Genome-wide association analyses in 128, 266 individuals identifies new morningness and sleep duration loci. *PLoS Genet* 2016;**12**:e1006125.
25. Perry JR, Day F, Elks CE *et al.* Parent-of-origin-specific allelic associations among 106 genomic loci for age at menarche. *Nature* 2014;**514**:92–97.
26. de Moor MH, Costa PT, Terracciano A *et al.* Meta-analysis of genome-wide association studies for personality. *Mol Psychiatry* 2012;**17**:337–49.
27. Lambert JC, Ibrahim-Verbaas CA, Harold D *et al.* Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat Genet* 2013;**45**:1452–58.
28. Burgess S, Thompson SG. Use of allele scores as instrumental variables for Mendelian randomization. *Int J Epidemiol* 2013; **42**:1134–44.
29. Euesden J, Lewis CM, O'Reilly PF. PRSice: polygenic risk score software. *Bioinformatics* 2015;**31**:1466–68.
30. StataCorp. *Stata Statistical Software: Release 14*. College Station, TX: StataCorp LP, 2015.
31. Marchini J, Howie B, Myers S, McVean G, Donnelly P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet* 2007;**39**:906–13.
32. Dudbridge F, Gusnanto A. Estimation of significance thresholds for genomewide association scans. *Genet Epidemiol* 2008;**32**: 227–34.
33. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* 2011;**88**: 76–82.
34. Pruim RJ, Welch RP, Sanna S *et al.* LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics* 2010;**26**:2336–37.
35. Davey Smith G, Ebrahim S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *Int J Epidemiol* 2003;**32**:1–22.
36. Vassos E, Di Forti M, Coleman J *et al.* An examination of polygenic score risk prediction in individuals with first-episode psychosis. *Biol Psychiatry* 2017;**81**:470–77.
37. Palla L, Dudbridge F. A fast method that uses polygenic scores to estimate the variance explained by genome-wide marker panels and the proportion of variants affecting a trait. *Am J Hum Genet* 2015;**97**:250–59.
38. Seaman SR, White IR. Review of inverse probability weighting for dealing with missing data. *Stat Methods Med Res* 2013;**22**: 278–95.
39. Lipsitch M, Tchetgen Tchetgen E, Cohen T. Negative controls: a tool for detecting confounding and bias in observational studies. *Epidemiology* 2010;**21**:383–88.
40. Xie PX, Kranzler HR, Yang C, Zhao HY, Farrer LA, Gelernter J. Genome-wide association study identifies new susceptibility loci for posttraumatic stress disorder. *Biol Psychiatry* 2013;**74**: 656–63.
41. Duncan LE, Ratanatharathorn A, Aiello AE *et al.* Largest GWAS of PTSD (N=20 070) yields genetic overlap with schizophrenia and sex differences in heritability. *Mol Psychiatry* 2018;**23**: 666–73.
42. Golias C, Charalabopoulos A, Stagikas D, Charalabopoulos K, Batistatou A. The kinin system—bradykinin: biological effects and clinical implications. Multiple role of the kinin system—bradykinin. *Hippokratia* 2007;**11**:124–28.
43. Burdett T, Hall PN, Hasting E *et al.* The NHGRI-EBI Catalog of published genome-wide association studies. Available at: www.ebi.ac.uk/gwas
44. Gage SH, Munafo MR, Davey Smith G. Causal inference in developmental origins of health and disease (DOHaD) research. *Annu Rev Psychol* 2016;**67**:567–85.
45. Goldberg M, Chastang JF, Leclerc A *et al.* Socioeconomic, demographic, occupational, and health factors associated with participation in a long-term epidemiologic survey: a prospective study of the French GAZEL cohort and its target population. *Am J Epidemiol* 2001;**154**:373–84.