Can education be personalised using pupils' genetic data?

Tim T Morris[*1,2], Neil M Davies[1,2], and George Davey Smith[1,2]

[1]MRC Integrative Epidemiology Unit at the University of Bristol, BS8 2BN, United Kingdom.

[2]Population Health Sciences, Bristol Medical School, University of Bristol, Barley House, Oakfield Grove, Bristol, BS8 2BN, United Kingdom.

[*]Corresponding author: Tim T Morris, MRC Integrative Epidemiology Unit, University of Bristol, Oakfield House, Oakfield Grove, BS8 2BN, UK. Email: Tim.Morris@bristol.ac.uk.

Key words: Genetics, education, prediction, ALSPAC.

Abstract

The predictive power of polygenic scores for some traits now rivals that of more classical phenotypic measures, and as such they have been promoted as a potential tool for genetically informed policy. However, how predictive polygenic scores are conditional on other easily available phenotypic data is not well understood. Using data from a UK cohort study, the Avon Longitudinal Study of Parents and Children, we investigated how well polygenic scores for education predict individuals' realised attainment over and above phenotypic data available to schools. Across our sample children's polygenic scores predicted their educational outcomes almost as well as parent's socioeconomic position or education. There was high overlap between the polygenic score and attainment distributions, leading to weak predictive accuracy at the individual level. Furthermore, conditional on prior attainment the polygenic score was not predictive of later attainment. Our results suggest that polygenic scores are informative for identifying group level differences, but they currently have limited use in predicting individual attainment.

Introduction

The increase in genetic discoveries from genomewide association studies (GWAS) has greatly advanced scientific understanding of the way in which complex social and health outcomes are influenced. GWAS with sample sizes of over one million participants have identified hundreds of genetic variants that associate with educational attainment and other social phenotypes[1–3]. While individual SNPs associate only very weakly with complex polygenic phenotypes in isolation, typically accounting for less than 0.01% of variation, together they can explain a considerable proportion of the phenotypic variation. For example, in the most recent education GWAS the median per allele effect size of lead variants related to an additional 1.7 weeks of schooling but all identified variants together explain 13% of the variance in years of education[1]. The combination of multiple variants in polygenic scores[4] - measures that sum the estimated effects of all individual genetic variants associated with a phenotype - are increasingly being used as indicators of genetic propensity, and have been promoted as a potential tool for genetically informed policy[5,6]. It has been suggested that genetic information could be used prescriptively to provide individually tailored medicine, education and even dating[5,7].

Claims have been made for and against the use of genetic data in personalising education[8–10]. For example, a proposed benefit is the potential to identify pupils in need of greater educational support[10], while a potential misuse could be the profiling of certain pupils based on genotype[9]. Polygenic scores constructed using a GWAS of over 1.1m individuals explained 11-13% of the variation in years of educational attainment in prediction samples[1]. Given the social complexity of educational attainment, these genetic scores associate with many aspects of environment and schooling[11,12], referred to as gene-environment correlation. This represents both active gene-environment correlation, where environment can be thought of as *down*stream of genotype (for example pupil's selecting certain subjects), and passive gene-environment correlation, where environment can be thought of as *up*stream of genotype (for example, children of highly educated parents being more likely to inherit education associated environments as well as education associated genes[13]). That a person's education polygenic score associates with a range of phenotypic differences very early in life demonstrates the broad range of information captured by polygenic scores.

The theoretical maximum bounds placed on the predictive ability of polygenic scores have been discussed in detail elsewhere (see [14–16]). Briefly, polygenic scores are more predictive when genetic factors play a larger role in a phenotype (as measured by heritability) and in the case of binary phenotypes where prevalence in the outcome is higher[14–16]. It has been suggested that for complex genetic phenotypes, accurate prediction at the individual level may require a polygenic score that explains around 75% of the total genetic variance of the phenotype[15]. It is therefore possible that polygenic scores for education will require greater explanatory power for accurate individual prediction. For polygenic scores to be informative for personalised education and provide actionable information to inform effective policy, the scores must not only explain sufficient variation in educational attainment, but they must also explain sufficient variation over and above other readily available phenotypic data. Phenotypic measures that are predictive of educational attainment such as sex, month of birth and prior attainment[17,18] are readily available to schools while others measures such as parental education and socioeconomic position[19,20] are easily and cheaply collected. However, to date, few studies have investigated how well polygenic scores predict

individual level educational attainment conditional on observable phenotypes that are easily available to educators. Here we investigate what information pupils' genetics confers over prior attainment or other phenotypic characteristics.

In this paper we combine educational and genetic data from a UK cohort, the Avon Longitudinal Study of Parents and Children (ALSPAC), to investigate the use of genotypic data in predicting pupil attainment and their potential for personalised education. We answer three related questions: 1) How predictive of realised educational attainment are polygenic scores? 2) Does polygenic prediction outperform phenotypic prediction from family background measures available to schools? 3) What incremental increase in predictive performance do polygenic scores offer over and above phenotypic information?

Results

Group level polygenic score prediction

We created two polygenic scores for education based on the results of the latest GWAS for educational attainment, the first using SNPs that reached genomewide significance and the second using all SNPs. Educational attainment measured as fine graded point scores correlated higher with the all SNP polygenic score (r for age 16 = 0.368) than the genomewide significant polygenic score (r for age 16 = 0.190) (Table 1). Children with higher polygenic scores, on average, had higher exam scores than those with lower polygenic scores. Correlations were similar between attainment and parents' years of education and highest parental socioeconomic position. Correlations were consistently stronger for age 16 than age 7 educational attainment.

Table 1: Correlation coefficients between educational attainment at ages 7 and 16 and the genotypic and social predictors. EA, educational attainment. PGS, polygenic score.

	EA 7	EA 16
GWAS sig PGS	0.167	0.190
All SNP PGS	0.264	0.368
Mothers EA	0.275	0.387
Fathers EA	0.271	0.397
Parents SEP	0.297	0.408

Figure 1 displays the explanatory power (Panel A) and incremental gain (Panel B) in variance explained of educational attainment at age 7 from the polygenic scores. Age 7 represents the first stage of assessment in UK schooling. The polygenic scores accounted for a larger amount of variation in attainment than age and sex alone, and when combined with attainment indicators available to schools explained up to 21.5% (95% CI: 18.9 to 24.1) of the variation in age 7 attainment. Including information on the social background of pupils' parents that is potentially obtainable by schools further increased the explanatory power of the models up to a maximum R2 of 26.3% (23.5 to 29.1). The incremental R2 of the polygenic scores over the information available to schools were 1.8% (-0.7 to 4.3) and 4.8 (2.2 to 7.3), suggesting that they provide additional predictive information over currently collected data.

The polygenic scores were more predictive of attainment in exams sat at the end of compulsory education at age 16 (Figure 2), explaining 6.6% (4.9 to 8.3) and 16.2% (13.8 to 18.5) of educational attainment at age 16 in our sample. This represents an incremental increase in explanatory power of 3.4% (1.7 to 5.0) and 12.9% (10.6 to 15.3) over age and sex alone; far higher returns than for age 7 attainment. At this stage of education schools also have data on pupils' prior attainment, and these measures explained a large amount of variation in age 16 attainment. For example, prior attainment at age 14 explained 61.9% (57.6 to 66.2) of the variation in age 16 attainment. Conditional on prior attainment data, the polygenic scores provide very little discernible increase in predictive power. By comparison, measures of parental education and socioeconomic position provided greater returns to predictive power than the polygenic scores when unadjusted for prior attainment, explaining an additional 19.1% (16.7 to 21.5) and 21.4% (18.9 to 24.0) respectively over age and sex. Again however, genotype and social background combined explained the largest amount of variation.

Figure 1: Variance in age 7 educational attainment explained by the polygenic scores. Educational attainment measured as exam point scores at age 7. Polygenic scores (PGS) created from SNPs associated with EA at p<5x10-8 (GWAS sig SNPs) and all SNPs (All SNP PGS). Attainment indicators include Free School Meals (FSM), English as a Foreign language (EFL) and Special Educational Needs (SEN) status taken from school records. Parents educational attainment measured as years of completed education. Parents socioeconomic position (SEP) measured as highest parental score on the CAMSIS social class scale. All analyses include adjustment for the first 20 principal components of population stratification. Parameter estimates in Supplementary Tables 1 and 2.

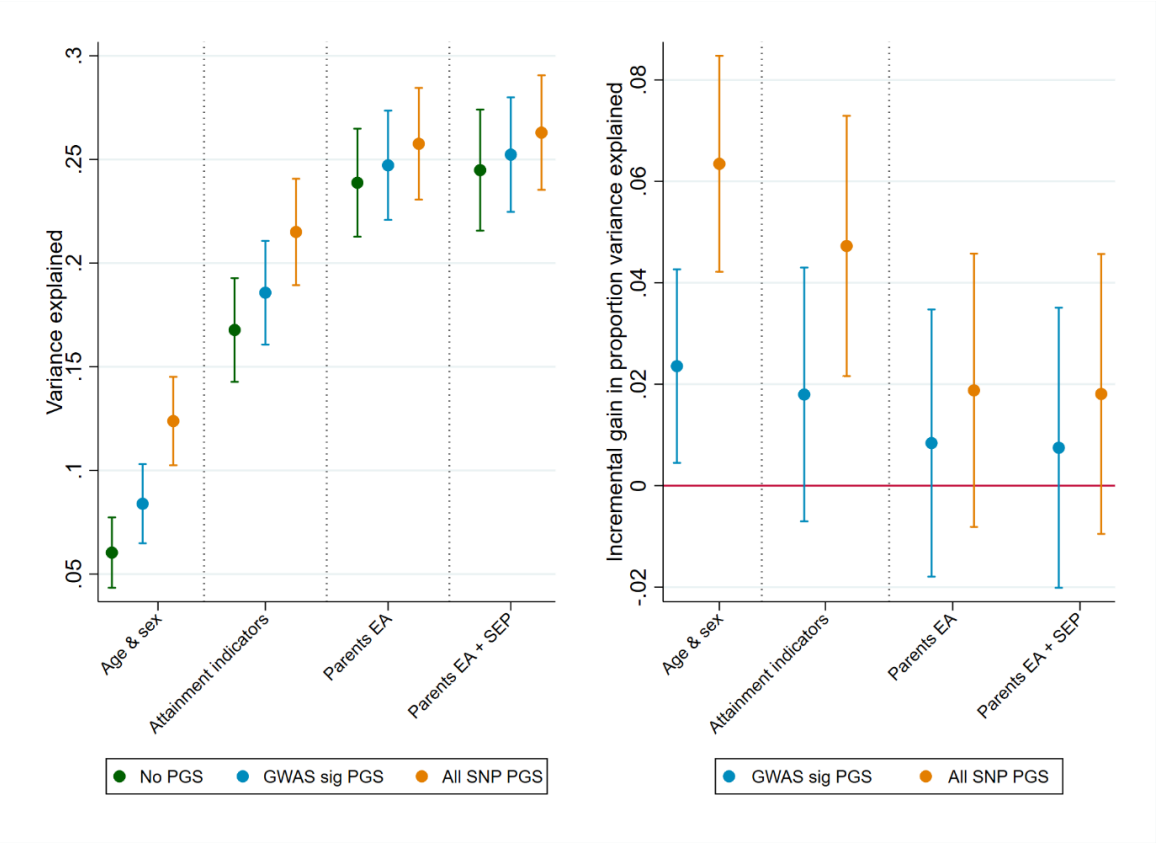
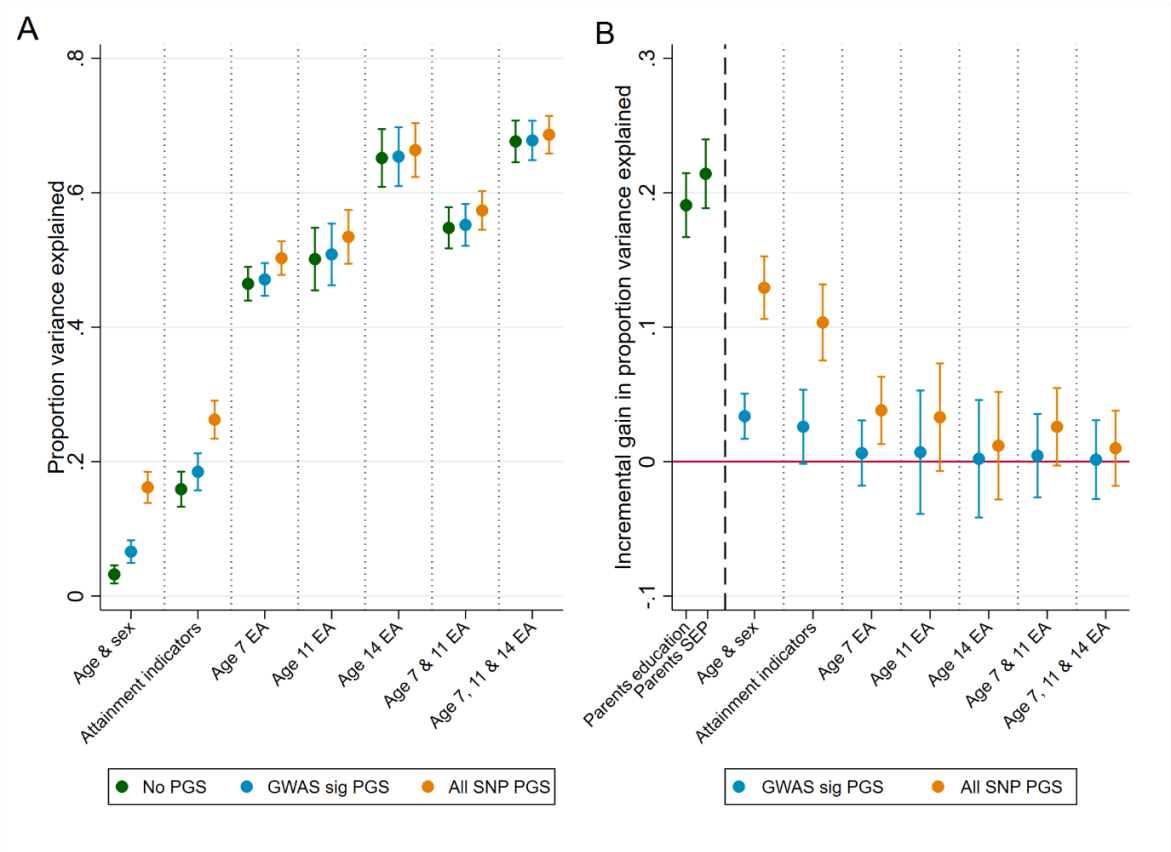


Figure 2: Variance in age 16 attainment explained by the polygenic scores. Educational attainment (EA) measured as exam point scores at ages 7, 11, 14 and 16. Polygenic scores (PGS) created from SNPs associated with EA at p<5x10-8 (GWAS sig SNPs) and all SNPs (All SNP PGS). Attainment indicators include Free School Meals (FSM), English as a Foreign language (EFL) and Special Educational Needs (SEN) status taken from school records. Parents educational attainment measured as years of completed education. Parents socioeconomic position (SEP) measured as highest parental score on the CAMSIS social class scale. All analyses include adjustment for the first 20 principal components of population stratification. Parameter estimates in Supplementary Tables 3 and 4.



Individual level polygenic score prediction

We next investigated how well the polygenic scores could identify high attaining pupils by splitting attainment and the polygenic scores by quintiles and deciles. The agreement between the quantiled measures of attainment at ages 7 and 16 and the polygenic scores were slightly higher than expected (Table 2). For the GWAS significant polygenic score, the Kappa statistics show that agreement was at most only 5% higher than would be expected by random agreement compared to perfect agreement (quintiles κ = 0.05 for attainment at age 7). Agreement was higher for the all SNP polygenic score and generally higher for age 16 than age 7 attainment. At age 16, agreement was at least twice as high for quantiles of prior attainment than the polygenic scores when compared to random allocation. Agreement with age 16 attainment was highest for age 14 attainment, being 46% and 28% better than expected by chance for quintiles and deciles of attainment respectively.

Table 2: Agreement between educational attainment quantiles and other quantiled measures. EA, educational attainment. PGS, polygenic score. p values for all tests <0.001.

	Age 7 attainment		Age 16 attainment	
	Agreement	Kappa	Agreement	Kappa
Quintiles				
Genotypic predictor				
GWAS sig PGS	24.1%	0.05 (0.01)	23.2%	0.04 (0.01)
All SNPs PGS	26.3%	0.08 (0.01)	28.2%	0.10 (0.01)
Phenotypic predictor				
Age 14 EA			56.5%	0.46 (0.01)
Age 11 EA			46.8%	0.33 (0.01)
Age 7 EA			37.8%	0.22 (0.01)
Deciles				
Genotypic predictor				
GWAS sig PGS	11.4%	0.02 (0.01)	12.4%	0.03 (0.01)
All SNPs PGS	13.9%	0.04 (0.01)	14.3%	0.05 (0.01)
Phenotypic predictor				
Age 14 EA			35.1%	0.28 (0.01)
Age 11 EA			27.5%	0.19 (0.01)
Age 7 EA			20.9%	0.12 (0.01)

Figure 3 shows the distributions of the two polygenic scores for pupils who were in the top 10% of age 16 attainers and all other pupils. The polygenic scores of high attainers are - on average - higher than of other pupils, but there is near complete overlap in the polygenic score distributions between the groups. This suggests there would be a large proportion of false positives when trying to predict from genetic data whether a pupil will be in the top 10%. By comparison, there is far less overlap in the distributions of prior attainment between high attainers and other pupils (Supplementary Figure S1). Figure 4 displays the misclassification of pupils graphically based on pupils being in the top 10% of attainers and the top 10% of genetically predicted attainers.

Figure 3: Distributions of polygenic scores between "high attainers" and all other pupils. High attainers defined as pupils with age 16 educational exam scores in the top 10% of the sample. Polygenic scores created from SNPs associated with educational attainment at p<5x10-8 (GWAS significant SNP polygenic score) and all SNPs (All SNP polygenic score).

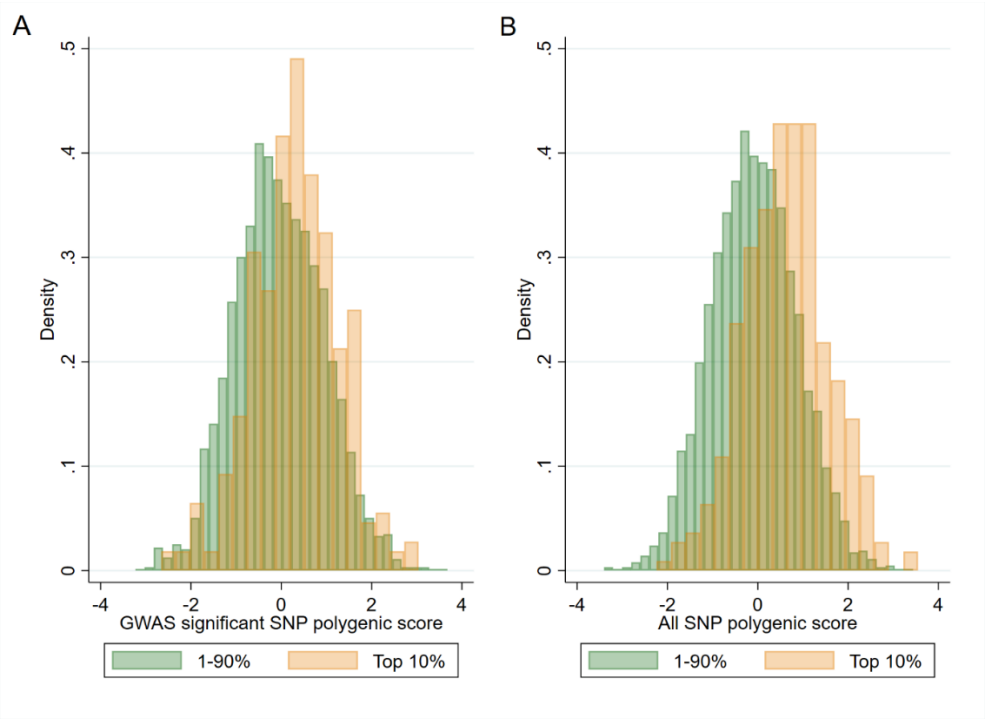
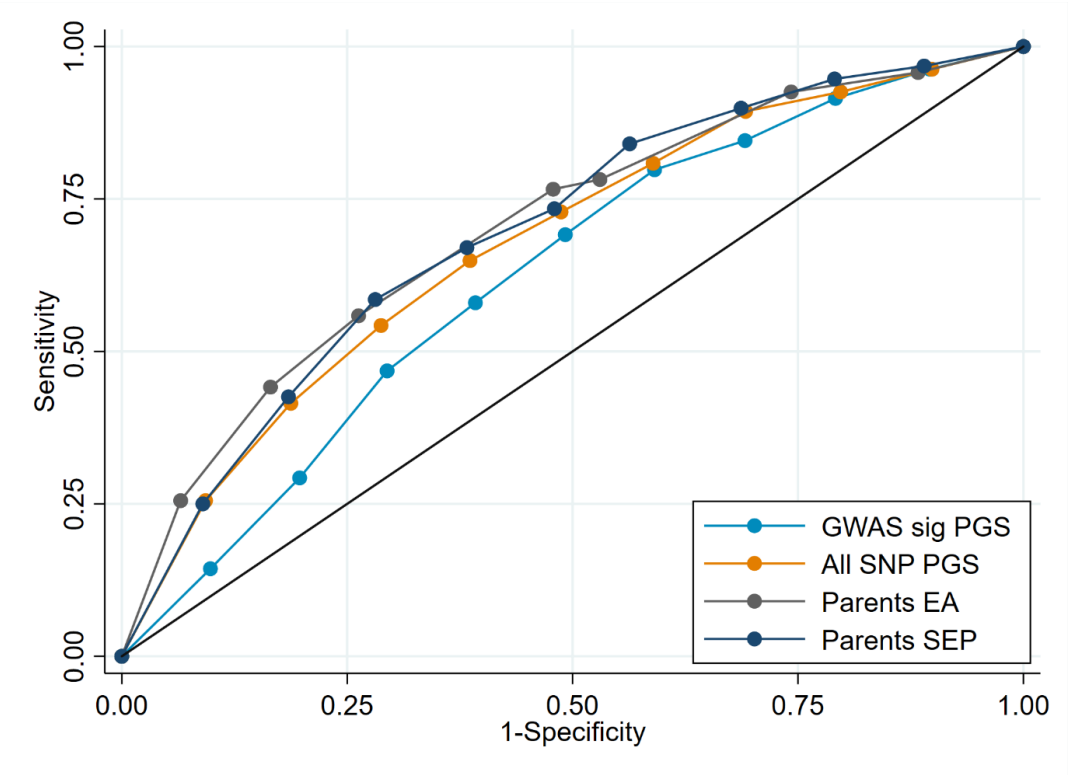


Figure 4: Correlation between realised and genetically predicted attainment. Educational attainment measured as exam point score at age 16. Predicted attainment at age 16 generated from the all SNP polygenic score. Solid lines separate pupils above and below the top decile of educational attainment at age 16 (high attainers) on the y axis and the top decile of those predicted to be in the top decile of educational attainment at age 16 from genetic data on the x axis. Dotted line represents best fit.

To investigate the potential performance of polygenic scores for correctly identifying high attaining students from all other pupils, we used Receiver Operating Characteristic (ROC) curves to calculate Area Under the Curve (AUC). ROC curves assess the sensitivity (the true positive rate, in our case the probability that a high attaining pupil will be correctly identified as a high attainer) and the specificity (the true negative rate, in our case the probability that that all other pupils will be correctly identified as not being high attainers) of a classifier as its discrimination threshold is varied. Compared to measures of parental socioeconomic position (AUCs: 0.70 for both years of education and social class), the polygenic scores have a lower AUC and therefore poorer sensitivity and specificity to discriminate high attainers at age 7 (AUCs: 0.63 for the GWAS PGS; 0.68 for the all SNP PGS) (Figure 5). The trade-off in sensitivity and specificity for each of the measures at different classification thresholds is also poor; high sensitivity comes at the cost of low specificity (and vice versa). This means that in order to set classification at the point where almost all students who go on to be in the top 10% of attainers are correctly identified, most students who don't go on to be in the top 10% would *also* be classified as high attaining students.

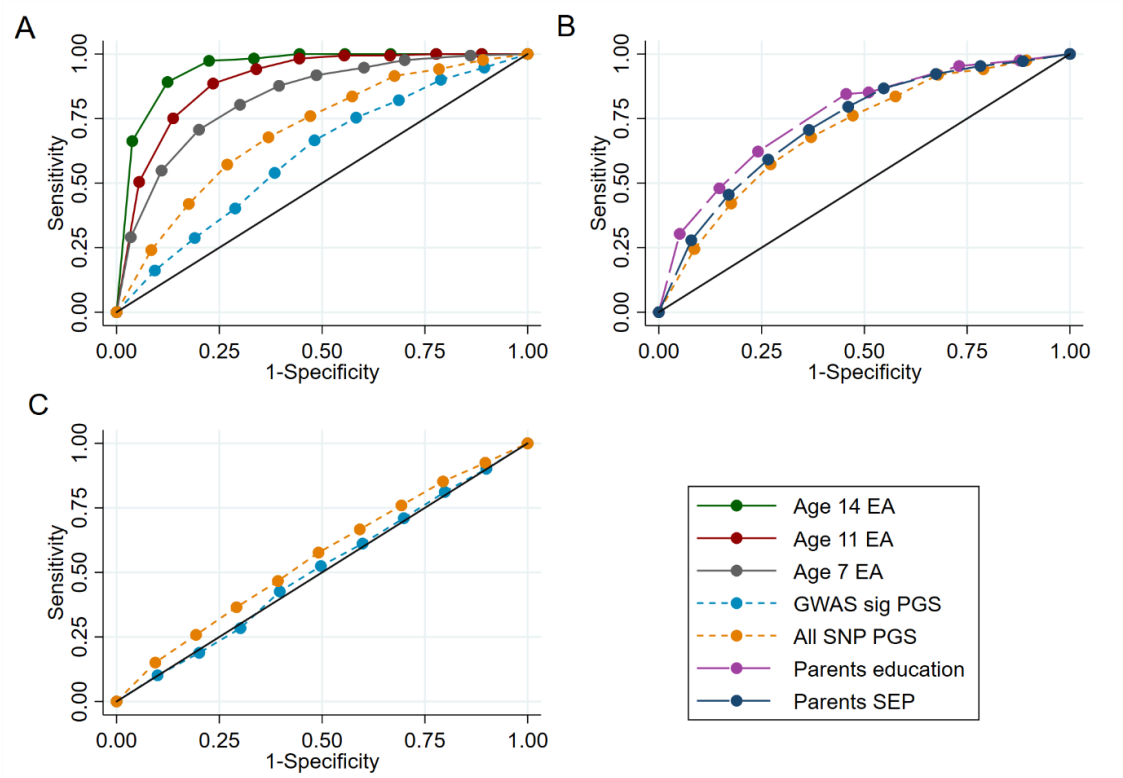
Figure 5: ROC curve for being a high attainment student (defined as the top 10% of pupils) at age 7. Parents educational attainment (EA) measured as years of completed education. Parents socioeconomic position (SEP) measured as highest parental score on the CAMSIS social class scale. Polygenic scores (PGS) created from SNPs associated with EA at p<5x10-8 (GWAS sig SNPs) and all SNPs (All SNP PGS). All PGS analyses include adjustment for the first 20 principal components of population stratification. Note that x axis displays 1-specificty.



For educational attainment at age 16 when prior attainment data are available, Figure 6a displays that these measures provide far higher sensitivity and specificity than the polygenic scores (AUCs: 0.83 to 0.95 for prior attainment compared to 0.61 to 0.70 for the polygenic scores). That is, there is a far better trade-off between sensitivity and specificity for prior attainment at age 14 than for either polygenic score. For example, a classification point can be set for prior attainment at age 14

where roughly 85% of students in both groups are accurately identified. For the polygenic scores, the best classification point would result in roughly two thirds of students being misclassified in both groups. As with attainment at age 7, the ROC curve for the all SNP polygenic score was similar to the ROC curves for parent's years of education and socioeconomic position (Figure 6b). We also calculated conditional ROC curves to test how well the polygenic scores identified pupils in the top decile of residualised age 16 attainment (conditional on age, sex, prior attainment, and indicators of low attainment). These demonstrated that after accounting for the phenotypic information available to schools, the polygenic scores do not provide any clear incremental predictive power for identifying high attainers (Figure 6c).

Figure 6: ROC curves for being a high attainment student (defined as the top 10% of pupils) at age 16. Panel A: Independent ROC curves for deciled measures of prior attainment and polygenic scores (PGS) predicting high educational attainment (EA) at age 16. Panel B: Independent ROC curves for deciled measures of parental education and socioeconomic position predicting high educational attainment (EA) at age 16. Panel C: Conditional ROC curves for deciled polygenic scores. Parents educational attainment (EA) measured as years of completed education. Parents socioeconomic position (SEP) measured as highest parental score on the CAMSIS social class scale. Polygenic scores (PGS) created from SNPs associated with EA at p<5x10-8 (GWAS sig SNPs) and all SNPs (All SNP PGS). All PGS analyses include adjustment for the first 20 principal components of population stratification. Note that x axis displays 1-specificty.



Discussion

We investigated how predictive polygenic scores for education were of realised attainment and the incremental increase in predictive power that they offered over and above readily available phenotypic measures. Our results demonstrated that the polygenic scores were predictive of educational attainment, accounting for 6.6% and 16.2% of variance in educational attainment at age 7 and 16 respectively. This is higher than the 2.5% to 13.8% reported in the original GWAS[1], and may reflect that we use fine points scores in exam results instead of years of completed education as our outcome measure. For informative education predictions at the individual level, the most predictive measure was prior attainment. This reflects practice whereby pupils may be streamed into certain classes in schools based upon teacher ratings and prior attainment. Conditional on prior attainment there was little incremental gain in the predictive power of polygenic scores for subsequent attainment. Our findings therefore suggest that when prior attainment data are available, polygenic scores are of little utility to providing accurate predictions of a child's future attainment. When children start school and prior attainment data are unavailable though, the scores may provide a small amount of predictive power. The lack of genotypic predictive power over and above phenotypic data may be because prior attainment mediates the effects of the genotypes on educational outcomes; genetic variants that affect educational attainment at earlier ages are likely to also affect attainment at later ages. More powerful polygenic scores may allow for better prediction of educational attainment in the future. However, our polygenic scores were constructed using results from a GWAS of over a million people, meaning that far larger samples will be required. Furthermore, parents' socioeconomic position and education were more strongly predictive of attainment than a pupil's genome; genetic data from individuals provided little information on their future attainment over phenotypic data that is either available or easily obtainable by educators.

At the individual level, polygenic scores and parental social background provided similar, but relatively imprecise predictions of attainment within our sample. This reflects a wider issue of the different challenges in analysing group and individual level differences[21]: while stochastic events will be averaged out at the group level, they are important in determining outcomes at the individual level. There was a large amount of overlap in the polygenic score distribution between pupils in the top 10% of attainers and all others; while pupils with a high polygenic score are more likely to be high attainers, genetics do not determine high attainment. High academic attainment is due to both environmental and genetic factors, including social background[19], teacher bias[22,23], the home and school environment[24,25], and luck. It is also possible that the quality of family and school environments may constrain or support pupils' ability to exploit their genetic propensity to education. For example, without the means to attend university, it does not matter what an individual's genotype is. In this, it is the combination of nature and nurture that is important. Such gene environment interactions have previously been demonstrated for education[26,27].

The usefulness of polygenic scores for educational research has been previously demonstrated for example in assessing the effectiveness of teachers and schools[12,23]; selection differences between schools[28,29]; and social mobility over time and space[30]. However, our results demonstrate that while polygenic scores are useful for investigating group differences, they do not provide suitable value for routine use by teachers and schools above phenotypic data to predict a pupil's attainment. Even at the current price of £50 per individual, the cost to genotype all 653,715 children who entered UK primary school in the 2017/2018 academic year would be over £32 million. While this would be

transformative for educational researchers, it is doubtful from these results whether it would provide useful additional information about pupils to teachers. There is a wide range of non-genetic information available to teachers as part of their day to day interactions with pupils that they use to inform and personalise their teaching. This may include knowledge of what the pupil responds well to, any stressful life events that they have recently experienced, and their physical and mental health. To the extent that this knowledge captures genotypic information of the pupil (through its expression in phenotype), it is unclear what novel information genotype would offer to teachers.

This study has several limitations. First, the ALSPAC cohort is not fully representative of the UK population and as such our results may not be generalisable to all UK pupils. Other studies, such as the Millennium Cohort Study are more representative and therefore could provide further evidence about personalised education for the broader UK population. Second, the educational attainment polygenic score that we use was based on a GWAS of years of education rather than exam scores. Years of education can be considered a more social measure of education than exam performance, and previous work has demonstrated that the educational attainment polygenic score strongly reflects parental social position (and through this access to further or higher education)[31].Future research could investigate this possibility by conducting a GWAS on detailed standardized exam scores on a large sample. Third, while the educational attainment polygenic score accounts for around 13% of the variance in years of education in our data, increases to this from future meta-analyses will provide greater power. Twin studies have estimated that the heritability of educational attainment is around 40%[32], which limits the predictive power of genetic measures for education over some other phenotypes[33]. Fourth, issues from confounding biases caused by population stratification and family level effects such as assortative mating and dynastic effects (genetic nurture)[13,31,34] may have impacted our results. These biases can lead to social and family differences being masked as genetic differences between individuals, inflating associations between polygenic scores and educational attainment in between individual analyses. Family data are required to further investigate the impact of these baises[35].

In conclusion, our results suggest that currently available genetic scores are unlikely to provide additional information about how well a pupil will perform in school over and above more readily available and easily collected phenotypic data, except where prior attainment measures are unavailable. The greatest value of genetic data may lie instead for researchers investigating performance differences between groups of pupils, teachers and schools and for novel analyses into socioeconomic inequalities in education attainment.

Materials and methods

Study sample

Participants were children from the Avon Longitudinal Study of Parents and Children (ALSPAC). Pregnant women resident in Avon, UK with expected dates of delivery 1st April 1991 to 31st December 1992 were invited to take part in the study. The initial number of pregnancies enrolled was 14,541. When the oldest children were approximately 7 years of age, an attempt was made to bolster the initial sample with eligible cases who had failed to join the study originally. This additional recruitment resulted in a total sample of 15,454 pregnancies, resulting in 14,901 children who were alive at one year of age. From this sample genetic data was available for 7,988 after

quality control and removal of related individuals. For full details of the cohort profile and study design see [36,37]. Please note that the study website contains details of all the data that is available through a fully searchable data dictionary and variable search tool at http://www.bristol.ac.uk/alspac/researchers/our-data/. The ALSPAC cohort is largely representative of the UK population when compared with 1991 Census data; there is under representation of some ethnic minorities, single parent families, and those living in rented accommodation[36]. Ethical approval for the study was obtained from the ALSPAC Ethics and Law Committee and the Local Research Ethics Committees. Following listwise deletion of cases with missing data our final analytical sample was 3,453 (Supplementary Figure S2).

Genetic data

DNA of the ALSPAC children was extracted from blood, cell line and mouthwash samples, then genotyped using references panels and subjected to standard quality control approaches. ALSPAC children were genotyped using the Illumina HumanHap550 quad chip genotyping platforms by 23andme subcontracting the Wellcome Trust Sanger Institute, Cambridge, UK and the Laboratory Corporation of America, Burlington, NC, US. The resulting raw genome-wide data were subjected to standard quality control methods. Individuals were excluded on the basis of gender mismatches; minimal or excessive heterozygosity; disproportionate levels of individual missingness (>3%) and insufficient sample replication (< 0.8). Population stratification was assessed by multidimensional scaling analysis and compared with Hapmap II (release 22) European descent (CEU), Han Chinese, Japanese and Yoruba reference populations; all individuals with non-European ancestry were removed. SNPs with a minor allele frequency of < 1%, a call rate of < 95% or evidence for violations of Hardy-Weinberg equilibrium ($P < 5x10^{-7}$) were removed. Cryptic relatedness was measured as proportion of identity by descent (IBD) > 0.1. Related subjects that passed all other quality control thresholds were retained during subsequent phasing and imputation. 9,115 participants and 500,527 SNPs passed these quality control filters. ALSPAC mothers were genotyped using the Illumina human660W-quad array at Centre National de Génotypage (CNG) and genotypes were called with Illumina GenomeStudio. PLINK (v1.07) was used to carry out quality control measures on an initial set of 10,015 subjects and 557,124 directly genotyped SNPs. SNPs were removed if they displayed more than 5% missingness or a Hardy-Weinberg equilibrium P value of less than 1.0e-06. Additionally SNPs with a minor allele frequency of less than 1% were removed. Samples were excluded if they displayed more than 5% missingness, had indeterminate X chromosome heterozygosity or extreme autosomal heterozygosity. Samples showing evidence of population stratification were identified by multidimensional scaling of genome-wide identity by state pairwise distances using the four HapMap populations as a reference, and then excluded. Cryptic relatedness was assessed using an IBD estimate of more than 0.125 which is expected to correspond to roughly 12.5% alleles shared IBD or a relatedness at the first cousin level. Related subjects that passed all other quality control thresholds were retained during subsequent phasing and imputation. 9,048 subjects and 526,688 SNPs passed these quality control filters.

We combined 477,482 SNP genotypes in common between the sample of mothers and sample of children. We removed SNPs with genotype missingness above 1% due to poor quality (11,396 SNPs removed) and removed a further 321 subjects due to potential ID mismatches. This resulted in a dataset of 17,842 subjects containing 6,305 duos and 465,740 SNPs (112 were removed during liftover and 234 were out of HWE after combination). We estimated haplotypes using ShapeIT (v2.r644) which utilises relatedness during phasing. The phased haplotypes were then imputed to

the Haplotype Reference Consortium (HRCr1.1, 2016) panel of approximately 31,000 phased whole genomes. The HRC panel was phased using ShapeIt v2, and the imputation was performed using the Michigan imputation server. This gave 8,237 eligible children and 8,196 eligible mothers with available genotype data after exclusion of related subjects using cryptic relatedness measures described previously. Principal components were generated by extracting unrelated individuals (IBS < 0.05) and independent SNPs with long range LD regions removed, and then calculating using the `--pca` command in plink1.90.

Educational attainment

We use average fine graded point scores at four major Key Stages of education in the UK. These are Key Stage 1 (age 7), Key Stage 2 (age 11), Key Stage 3 (age 14), and Key Stage 4 (age 16). We use scores for performance at the end of each Key Stage and a score at entry to Key Stage 1, which represents the start of schooling. At the time the ALSPAC cohort were at school, the age 16 Key Stage 4 exams represented final compulsory schooling examinations. Scores were obtained through data linkage to the UK National Pupil Database (NPD), which represents the most accurate record of individual educational attainment available in the UK. We used data from the Key Stage 1 and Key Stage 4 files. The Key Stage 4 database provides a larger sample size than Key Stage 2 and 3 databases and contains data for each.

Educational attainment polygenic scores

Two educational attainment polygenic scores were generated using the software package PRSice[38] based upon the list of SNPs identified to associate with years of education in the largest GWAS of education to date[1]. The polygenic scores were generated using GWAS results which had removed ALSPAC and 23andMe participants from the meta-analysis. SNPs were weighted by their effect size in the replication cohort of the GWAS, and these sizes were summed using allelic scoring. PRSice was used to thin SNPs according to linkage disequilibrium through clumping, where the SNP with the smallest *P*-value in each 250kb window was retained and all other SNPs in linkage disequilibrium with an r^2 of >0.1 were removed. The first polygenic score (GWAS sig PGS) was created from the 1,271 independent SNPs that associated with years of education at genome-wide levels of significance ($p<5x10^{-8}$). The second (all SNP PGS) was created from all genome-wide SNPs reported in the meta-analysis.

Covariates

We selected covariates that are easily available to schools in the UK. These include the study participants sex and month of birth, and their status on three indicators of low attainment available in the NPD: eligibility for Free School Meals (FSM); Special Education Needs (SEN); and English as a foreign language (EFL). FSM is a proxy for low income as only children from low income families are eligible. We use years of parental education, coded as basic formal education (7 years), certificate of secondary education (10 years), O-levels and vocational qualifications (11 years), A-level (13 years), and degree (16 years). Finally, we use a continuous measure of socioeconomic position (SEP), the Cambridge Social Stratification Score (CAMSIS).

Statistical analysis

To examine the predictive ability of polygenic scores for educational attainment we ran a series of regression analyses of the polygenic scores on attainment each controlling for sex, month of birth, and the first 20 principal components of inferred population structure. Principal components are

included to adjust estimates for population stratification; systematic differences in allele frequencies between subpopulations due to ancestral differences. Predictive ability of the polygenic scores was determined by the incremental increase in variance explained (R^2) in educational attainment above age and sex; indicators of low attainment; and prior attainment. Bootstrapping with 1000 replications was used to estimate confidence intervals for R^2values. To compare the predictive power of polygenic scores to additional phenotypic data that schools could collect we repeated the regression analyses controlling for parental years of education, grandparental years of education and parental socioeconomic position. Sensitivity and specificity were calculated using selection into the top 10% of educational attainers at age 16 from the whole cohort as the 'diagnosis'. Receiver Operating Characteristic (ROC) curves were used to visually compare models and to calculate the Area Under the Curve (AUC).

References

1. Lee, J. J. *et al.* Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nat. Genet.* (2018). doi:10.1038/s41588-018-0147-3

2. Karlsson Linnér, R. *et al.* Genome-wide association analyses of risk tolerance and risky behaviors in over 1 million individuals identify hundreds of loci and shared genetic influences. *Nat. Genet.* (2019). doi:10.1038/s41588-018-0309-3

3. Luciano, M. *et al.* Association analysis in over 329,000 individuals identifies 116 independent variants influencing neuroticism. *Nat. Genet.* (2018). doi:10.1038/s41588-017-0013-8

4. Dudbridge, F. Power and Predictive Accuracy of Polygenic Risk Scores. *PLoS Genet.* **9**, (2013).

5. Plomin, R. *Blueprint: How DNA makes us who we are*. (Allen Lane, 2018).

6. Conley, D. C. & Fletcher, J. M. *The Genome Factor: What the Social Genomics Revolution Reveals about Ourselves, Our History, and the Future*. (Princeton University Press, 2017).

7. McCarthy, M. I. & Mahajan, A. The value of genetic risk scores in precision medicine for diabetes. *Expert Rev. Precis. Med. Drug Dev.* **3**, 279–281 (2018).

8. Miller, R. Beyond reductionism: The emerging holistic paradigm in education. *Humanist. Psychol.* (1990). doi:10.1080/08873267.1990.9976898

9. Grigorenko, E. L. How Can Genomics Inform Education? *Mind, Brain, Educ.* (2007). doi:10.1111/j.1751-228X.2007.00001.x

10. Sabatello, M. A genomically informed education system? Challenges for behavioral genetics. *Journal of Law, Medicine and Ethics* (2018). doi:10.1177/1073110518766027

11. Abdellaoui, A. *et al.* Genetic Consequences of Social Stratification in Great Britain. *bioRxiv* (2018). doi:10.1101/457515

12. Harden, K. P. *et al.* Genetic Associations with Mathematics Tracking and Persistence in Secondary School. *bioRxiv* (2019). doi:10.1101/598532

13. Kong, A. *et al.* The nature of nurture: Effects of parental genotypes. *Science (80-.).* **359**, 424–428 (2018).

14. Janssens, A. C. J. W. *et al.* Predictive testing for complex diseases using multiple genes: Fact or fiction? *Genet. Med.* (2006). doi:10.1097/01.gim.0000229689.18263.f4

15. Wray, N. R., Yang, J., Goddard, M. E. & Visscher, P. M. The genetic interpretation of area under the ROC curve in genomic profiling. *PLoS Genet.* (2010). doi:10.1371/journal.pgen.1000864

16. Zhao, B. & Zou, F. Is PRS Good for Predicting Complex Polygenic Traits? *bioRxiv* (2018). doi:10.1101/447797

17. Benton, T., Hutchison, D., Schagen, I. & Scott, E. *Study of the performance of maintained secondary schools in England*. (National Audit Office, 2004).

18. Solli, I. F. Left behind by birth month. *Educ. Econ.* **25**, 323–346 (2017).

19. Morris, T., Dorling, D. & Davey Smith, G. How well can we predict educational outcomes? Examining the roles of cognitive ability and social position in educational attainment. *Contemp. Soc. Sci.* **11**, 1–15 (2016).

20. Strand, S. The limits of social class in explaining ethnic gaps in educational attainment. *Br. Educ. Res. J.* **37**, 197–229 (2011).

21. Davey Smith, G. Epidemiology, epigenetics and the 'Gloomy Prospect': Embracing randomness in population health research and practice. *Int. J. Epidemiol.* **40**, 537–562 (2011).

22. Campbell, T. Stereotyped at Seven? Biases in Teacher Judgement of Pupils' Ability and Attainment. *J. Soc. Policy* **44**, 517–547 (2015).

23. Morris, T. T., Davies, N. M., Dorling, D., Richmond, R. C. & Davey Smith, G. Testing the validity of value-added measures of educational progress with genetic data. *Br. Educ. Res. J.* **0**, (2018).

24. Nieuwenhuis, J. & Hooimeijer, P. The association between neighbourhoods and educational achievement, a systematic review and meta-analysis. *J. Hous. Built Environ.* (2016). doi:10.1007/s10901-015-9460-7

25. Rasbash, J., Leckie, G., Pillinger, R. & Jenkins, J. Children's educational progress: Partitioning family, school and area effects. *J. R. Stat. Soc. Ser. A Stat. Soc.* **173**, 657–682 (2010).

26. de Zeeuw, E. L. & Boomsma, D. I. Country-by-genotype-by-environment interaction in childhood academic achievement. *Proc. Natl. Acad. Sci.* (2017). doi:10.1073/pnas.1718938115

27. Belsky, D. W. *et al.* Genetics and the geography of health, behaviour and attainment. *Nat. Hum. Behav.* (2019). doi:10.1038/s41562-019-0562-1

28. Smith-Woolley, E. *et al.* Differences in exam performance between pupils attending selective and non-selective schools mirror the genetic differences between them. *npj Sci. Learn.* (2018). doi:10.1038/s41539-018-0019-8

29. Trejo, S. *et al.* Schools as Moderators of Genetic Associations with Life Course Attainments: Evidence from the WLS and Add Heath. *Sociol. Sci.* (2018). doi:10.15195/v5.a22

30. Belsky, D. W. *et al.* Genetic analysis of social-class mobility in five longitudinal studies. *Proc. Natl. Acad. Sci.* (2018). doi:10.1073/pnas.1801238115

31. Bates, T. C. *et al.* The Nature of Nurture: Using a Virtual-Parent Design to Test Parenting Effects on Children's Educational Attainment in Genotyped Families. *Twin Res. Hum. Genet.* (2018). doi:10.1017/thg.2018.11

32. Branigan, A. R., Mccallum, K. J. & Freese, J. Variation in the heritability of educational attainment: An international meta-analysis. *Soc. Forces* **92**, 109–140 (2013).

33. Daetwyler, H. D., Villanueva, B. & Woolliams, J. A. Accuracy of predicting the genetic risk of disease using a genome-wide approach. *PLoS One* (2008). doi:10.1371/journal.pone.0003395

34. Young, A. I. *et al.* Relatedness disequilibrium regression estimates heritability without environmental bias. *Nat. Genet.* (2018). doi:10.1038/s41588-018-0178-9

35. Brumpton, B. *et al.* Within-family studies for Mendelian randomization: avoiding dynastic, assortative mating, and population stratification biases. *bioRxiv* (2019). doi:10.1101/602516

36. Boyd, A. *et al.* Cohort Profile: the 'children of the 90s'--the index offspring of the Avon Longitudinal Study of Parents and Children. *Int J Epidemiol* **42**, 111–127 (2013).

37. Fraser, A. *et al.* Cohort profile: The avon longitudinal study of parents and children: ALSPAC mothers cohort. *Int. J. Epidemiol.* **42**, 97–110 (2013).

38. Euesden, J., Lewis, C. M. & O'Reilly, P. F. PRSice: Polygenic Risk Score software. *Bioinformatics* **31**, 1466–1468 (2015).

Acknowledgements

We thank Kirsten Leyland for her extremely helpful comments on an earlier version of this manuscript. The Medical Research Council (MRC) and the University of Bristol support the MRC Integrative Epidemiology Unit [MC_UU_12013/1, MC_UU_12013/9, MC_UU_00011/1]. The Economics and Social Research Council (ESRC) support NMD via a Future Research Leaders grant [ES/N000757/1] and TTM via a postdoctoral fellowship [ES/S011021/1]. No funding body has influenced data collection, analysis or its interpretation. This publication is the work of the authors, who serve as the guarantors for the contents of this paper. 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. The UK Medical Research Council and Wellcome (Grant ref: 102215/2/13/2) and the University of Bristol provide core support for ALSPAC. A comprehensive list of grants funding is available on the ALSPAC website (http://www.bristol.ac.uk/alspac/external/documents/grant-acknowledgements.pdf). GWAS data was generated by Sample Logistics and Genotyping Facilities at Wellcome Sanger Institute and LabCorp (Laboratory Corporation of America) using support from 23andMe. No funding body has influenced data collection, analysis or its interpretations.

Competing interests

Neil Davies reports a grant for unrelated research from the Global Research Awards for Nicotine Dependence which is an Independent Competitive Grants Program supported by Pfizer.

Author contributions

TTM, NMD and GDS obtained funding for this study. TTM analysed and cleaned the data, interpreted results, wrote and revised the manuscript. NMD and GDS interpreted the results, wrote and revised the manuscript.