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, polygenic score, achievement, ALSPAC.

Abstract

The increasing predictive power of polygenic scores for education has led to their promotion by some as a potential tool for genetically informed policy. How well polygenic scores predict educational performance conditional on other phenotypic data is however not well understood. Using data from a UK cohort study, we investigated how well polygenic scores for education predicted pupils' realised achievement over and above phenotypic data that are available to schools. Across our sample, prediction of educational outcomes from polygenic scores were inferior to those from parental socioeconomic factors. There was high overlap between the polygenic score and achievement distributions, leading to weak predictive accuracy at the individual level. Furthermore, conditional on prior achievement polygenic scores were not predictive of later achievement. Our results suggest that while polygenic scores can be informative for identifying group level differences, they currently have limited use for predicting individual educational performance or for personalised education.

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 may arises. 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 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 explained up to 13% of the variance in years of education in prediction samples[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 personalised medicine, education and even dating[5,7].

Personalisation refers to the tailoring of services away from a one-size-fits-all model to a customised approach that focuses on the needs of an individual. The definition of personalised education has been inconsistent, generally referring to either the tailoring of educational curriculums, learning environments and teaching styles for individual students, or for groups of students within a classroom[8,9]. Personalised learning was adopted in national policy statements in England in 2004 with a focus on the needs of individual students[10,11]. However, it was not mandated and was seen as being conceptually ambiguous, leading to inconsistency in its implementation across schools[12]. Throughout, we refer to personalised education as administered at the individual level. There are currently no policies in place that rely on educational prediction, but calls are increasingly being made for genetic data to be used to personalise education[13–15]. For example, a proposed benefit is the potential to identify pupils in need of greater educational support[15]. Polygenic scores constructed using a GWAS of educational attainment (defined as completed years of education) in over 1.1m individuals explained up to 13% of the variation in attainment and 9.2% of the variation in achievement (defined as high school grade point average [GPA]) in US samples[1]. Given the social complexity of educational attainment, these genetic scores associate with many aspects of environment and schooling[16,17], referred to as gene-environment correlation. Active gene-environment correlation can be thought of as environment *down*stream of genotype; for example, pupil's selecting certain subjects based on their genotype. Passive gene-environment correlation can be thought of as environment *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[18] (also referred to as dynastic effects[19,20]). That a person's education polygenic score associates with a range of phenotypic differences very early in life demonstrates that polygenic scores capture a very broad range of information, not just their education.

The theoretical maximum bounds placed on the predictive ability of polygenic scores have been discussed in detail elsewhere (see [21–23]). 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[21–23]. 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 achievement (defined as performance in educational tests), but they must also explain sufficient variation over and above other readily available phenotypic data. Phenotypic measures that are predictive of educational achievement such as sex, month of birth and prior achievement[24,25] are readily available to schools while other measures such as parental education and socioeconomic position[26,27] are, in principle, simple and inexpensive to collect. To date, few studies have investigated how well polygenic scores predict individual level educational attainment or achievement conditional on observable phenotypes that are easily available to educators. Here we investigate what information pupils' genetics confers over prior achievement and 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 achievement and their potential for personalised education. We answer three related questions: 1) How predictive of realised educational achievement 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

To investigate how predictive polygenic scores are of realised educational achievement, we created two scores for education based on the results of the latest GWAS for educational attainment[1]. The first used SNPs that reached genomewide significance ($p<5x10^{-8}$) and the second used all education associated SNPs. Our measure of educational achievement was fine graded point scores from educational exams taken at ages 7 and 16. The all SNP polygenic score was more strongly correlated with educational achievement (r for age 16 = 0.37) than the genomewide significant polygenic score (r for age 16 = 0.19) (Table 1). Children with higher polygenic scores, on average, had higher exam scores than those with lower polygenic scores. Correlations were similar between achievement and parents' years of education and highest parental socioeconomic position. Correlations were consistently stronger for age 16 than age 7 educational achievement.

Table 1: Correlation coefficients between educational achievement at ages 7 and 16 and the genotypic and social predictors. Educational achievement measured using fine graded point scores from educational exams at ages 7 and 16. Genotypic predictors measured using two polygenic scores (PGS) built using only genome-wide significant SNPs (GWAS sig PGS) or all education associated SNPs (all SNP PGS) from the largest GWAS of educational attainment[1]. Parental educational attainment (EA) was measured as average completed years of education. Parental socioeconomic position (SEP) was measured as highest parental score on the Cambridge Social Stratification Score scale.

	Achievement age 7	Achievement 16
GWAS sig PGS	0.17	0.19
All SNP PGS	0.26	0.37
Mothers EA	0.28	0.39
Fathers EA	0.27	0.40
Parents SEP	0.30	0.40

Next, we assessed the predictive power of polygenic scores for educational achievement at age 7. We assessed this using the incremental gain in variance of educational achievement explained by the polygenic scores over and above pupil characteristics available to schools (age, sex, Free School Meal status, English as a Foreign Language status, Special Educational Needs status), parents years of education, and parents socioeconomic position (Figure 1). Both the genomewide significant and the all SNP polygenic scores accounted for a larger proportion of variance explained (R^2) in achievement than age and sex alone. Pupil characteristics outperformed polygenic scores in terms of explanatory power, but together they explained up to 21.5% (95% CI: 18.9 to 24.1) of the variation in age 7 achievement. 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 R^2 of 26.3% (23.4 to 29.2). The incremental R^2 of the polygenic scores over pupil characteristics were 1.8% (-0.7 to 4.3) and 4.8 (2.1 to 7.3), suggesting that they provide some additional predictive information over currently available or easily collectable data.

The genomewide significant and all SNP polygenic scores were more predictive of achievement in exams sat at the end of compulsory education at age 16, explaining an additional 3.4% (1.7 to 5.0) and 12.9% (10.6 to 15.3) of educational achievement over age and sex alone (Figure 2B). By comparison, measures of parental education and socioeconomic position provided greater returns to predictive power than the polygenic scores when unadjusted for prior achievement, explaining an additional 19% (16.6 to 21.4) and 21.4% (18.8 to 23.9) respectively over age and sex (Figure 2B). As with age 7 achievement, using both genotype and social background data explained the largest amount of variation. At this stage of education schools also hold data on pupils' prior achievement, and these prior achievement measures explained a large amount of variation in age 16 achievement. For example, prior achievement at age 14 explained 65.1% (60.9 to 69.4) of the variation in age 16 achievement alongside age and sex (Figure 2A). Conditional on prior achievement data, the polygenic scores provide very little discernible increase in predictive power (Figure 2B).

Figure 1: Variance in age 7 educational achievement explained by the polygenic scores. Educational achievement measured using fine graded point scores from educational exams at age 7. Polygenic scores (PGS) built using only genome-wide significant SNPs (GWAS sig SNPs) or all education associated SNPs (All SNP PGS) from the largest GWAS of educational attainment[1]. Pupil characteristics available to schools include Free School Meals (FSM), English as a Foreign language (EFL) and Special Educational Needs (SEN) status. Parental educational attainment was measured as average years of completed education. Parental socioeconomic position (SEP) was measured as highest parental score on the Cambridge Social Stratification Score scale. All analyses include adjustment for the first 20 principal components of population stratification. Parameter estimates in Supplementary Tables S2 and S3.

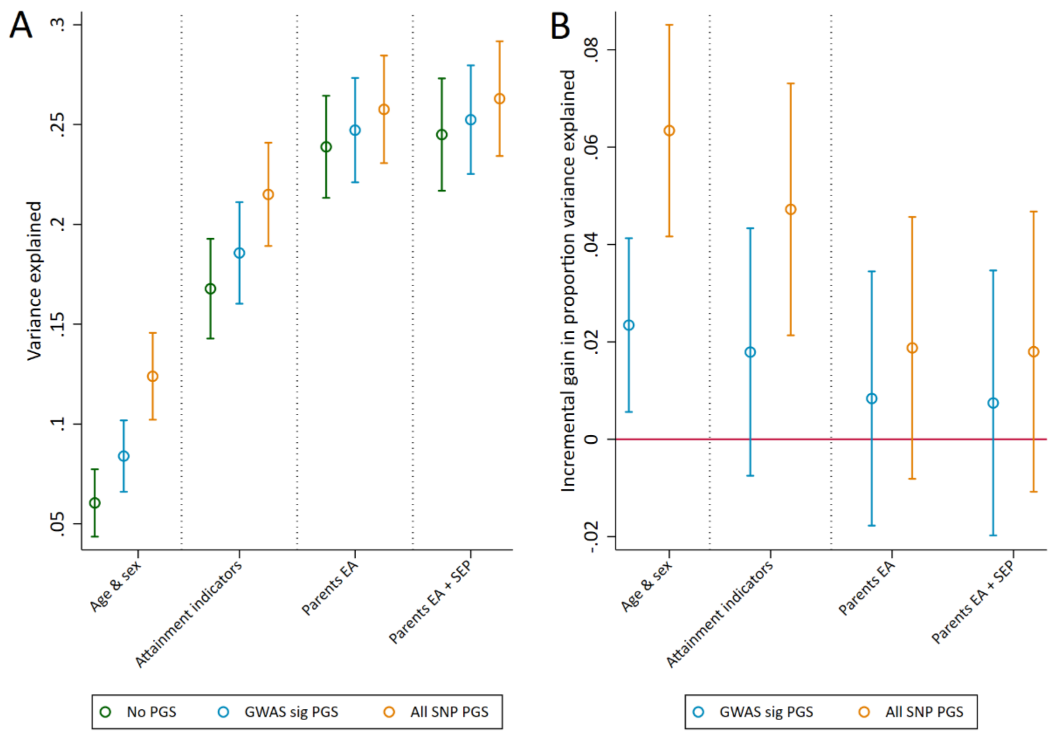
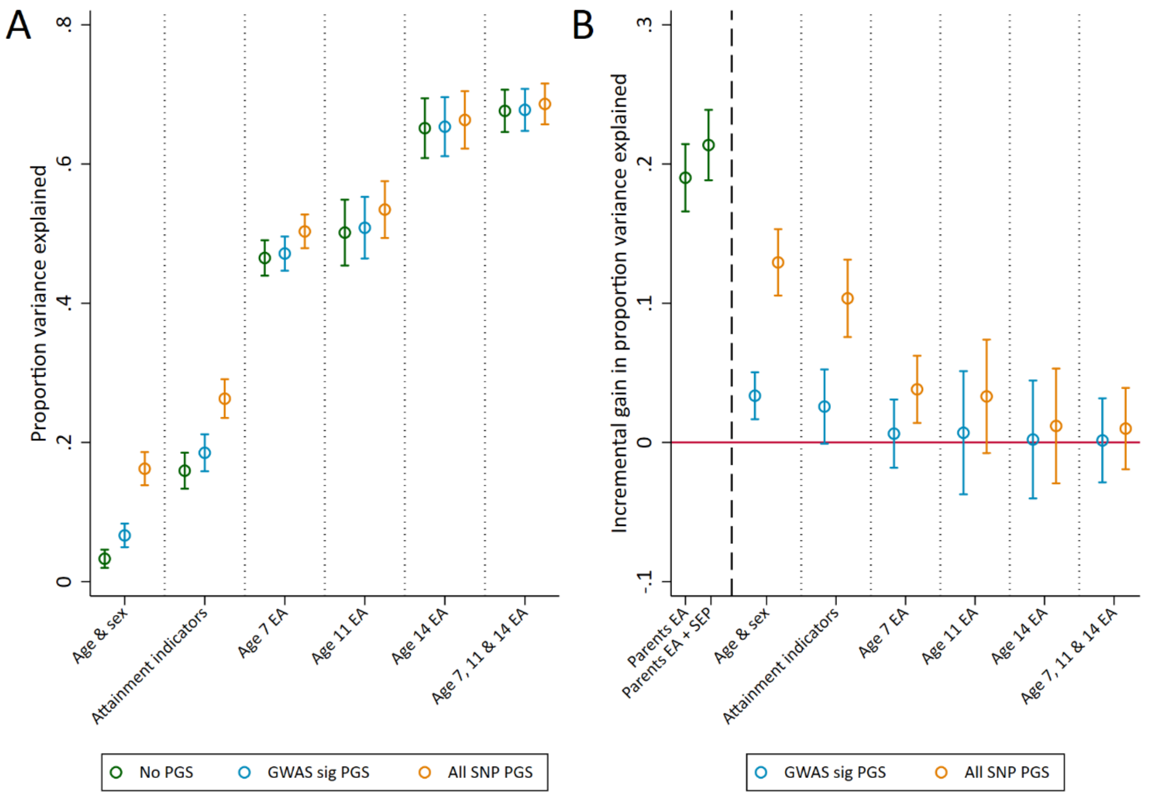


Figure 2: Variance in age 16 achievement explained by the polygenic scores. Educational achievement (EA) measured using fine graded point scores from educational exams at ages 7, 11, 14 and 16. Polygenic scores (PGS) built using only genome-wide significant SNPs (GWAS sig SNPs) or all education associated SNPs (All SNP PGS) from the largest GWAS of educational attainment[1]. Pupil characteristics available to schools include Free School Meals (FSM), English as a Foreign language (EFL) and Special Educational Needs (SEN) status. Parental educational attainment was measured as average years of completed education. Parental socioeconomic position (SEP) was measured as highest parental score on the Cambridge Social Stratification Score scale. All analyses include adjustment for the first 20 principal components of population stratification. Parameter estimates in Supplementary Tables S4 and S5.



Individual level polygenic score prediction

We next investigated how well the polygenic scores could identify high achieving pupils, defined as those with the highest 10% of educational test scores. Figure 3 shows the distributions of the two polygenic scores for high achieving pupils at age 16 and all other pupils. The polygenic scores of high achievers are - on average - higher than of other pupils, but there is near complete overlap in the distributions between the groups. This suggests there would be a large proportion of misclassification 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 achievement between high achievers and other pupils (Supplementary Figure S1). Figure 4 displays this misclassification of pupils; while some are correctly predicted from their genetic data to be high achievers, a greater proportion are erroneously predicted to be in the wrong group. This misclassification is similar for parental education and socioeconomic position but lower for prior attainment (Supplementary Figure S2). In each case, as a group the pupils predicted to be in the top 10% of achievers will on average perform higher than other pupils in exams, but the large variability shows that many of the pupils in this group will underperform. High levels of misclassification from the polygenic scores compared to prior attainment were also evident when assessing agreement with quantiled measures of educational achievement (Supplementary Table S1).

Figure 3: Distributions of polygenic scores between "high achievers" and all other pupils. High achievers defined as pupils with age 16 educational exam scores in the top 10% of the sample. Polygenic scores (PGS) built using only genome-wide significant SNPs (GWAS sig SNPs) or all education associated SNPs (All SNP PGS) from the largest GWAS of educational attainment[1].

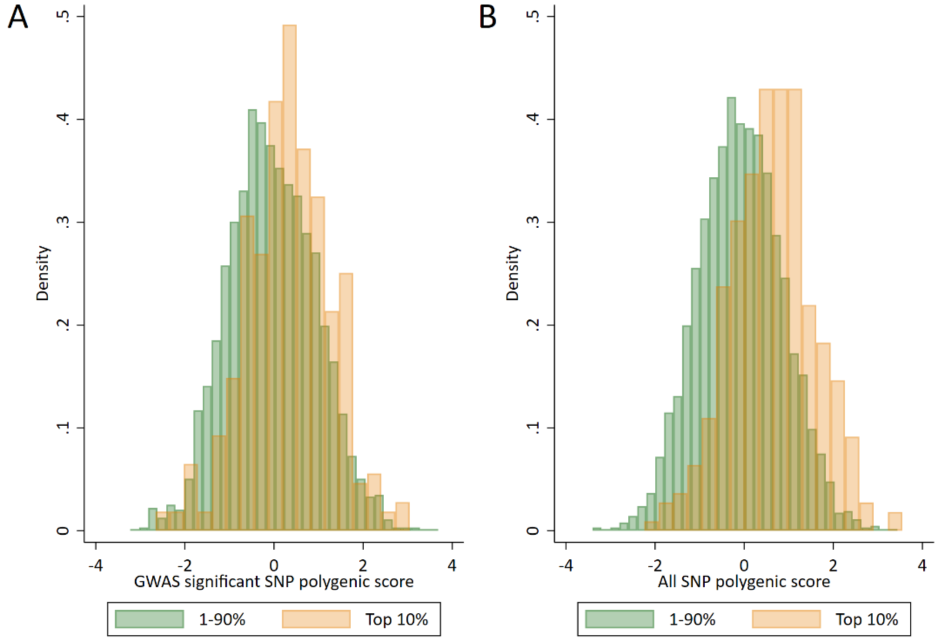
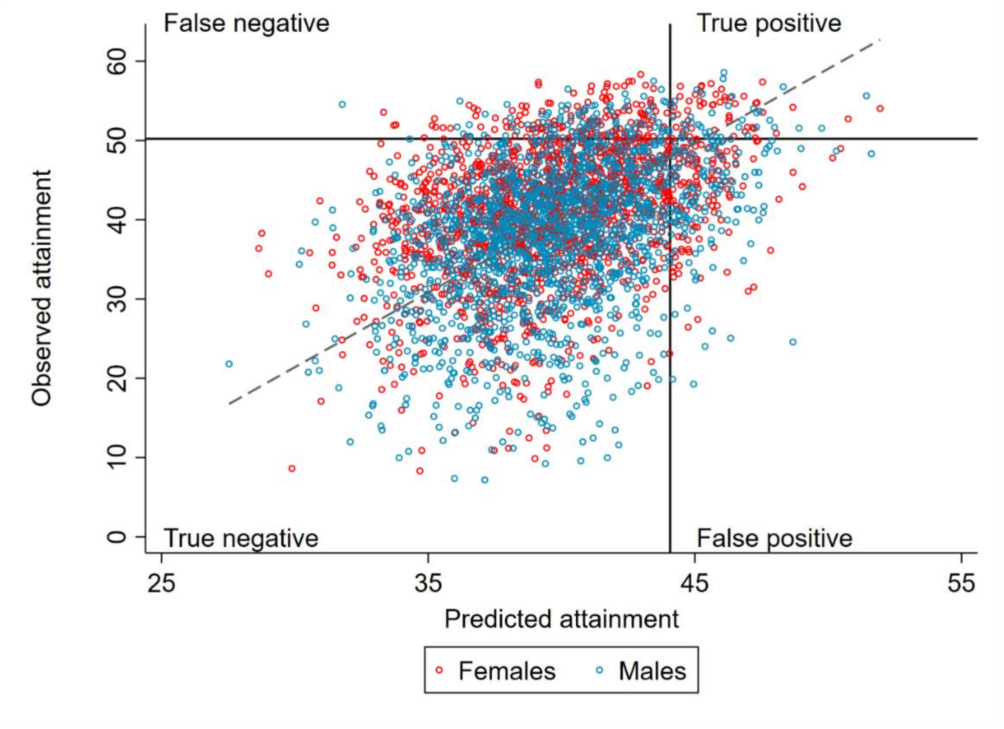


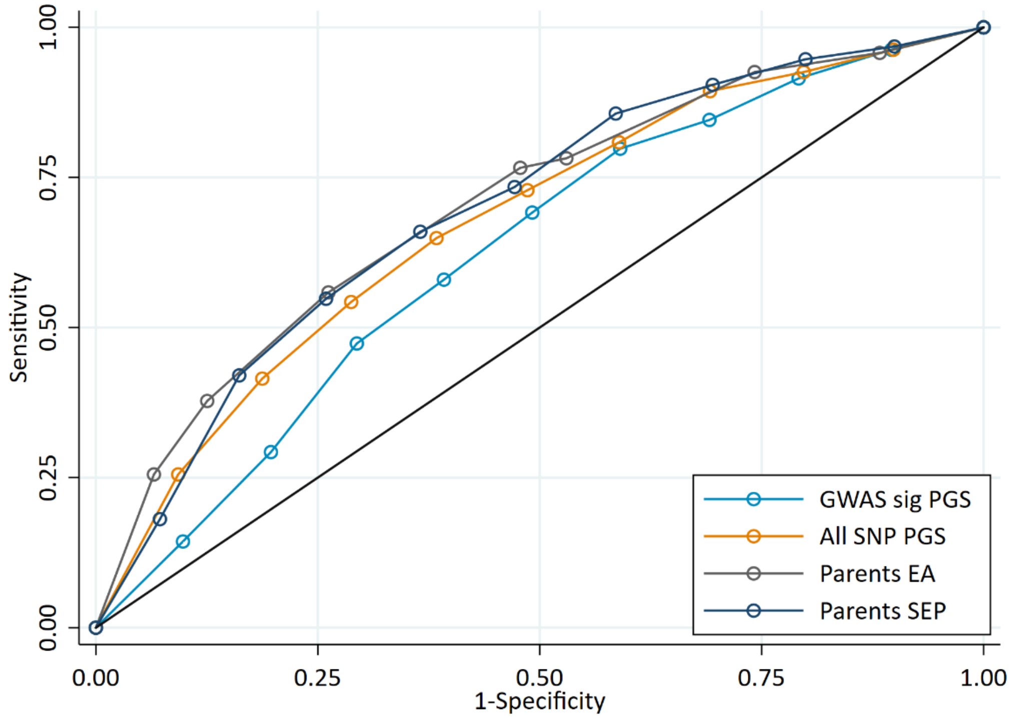
Figure 4: Correlation between realised and genetically predicted achievement. Educational achievement measured using fine graded point scores from educational exams at age 16. Predicted achievement at age 16 generated from a polygenic score built using all education associated SNPs (All SNP PGS) from the largest GWAS of educational attainment[1]. Solid lines separate pupils above and below the top decile of educational achievement at age 16 (high achievers) on the y axis and the top decile of those predicted to be in the top decile of educational achievement at age 16 from genetic data on the x axis. Dotted line represents best fit.



Using polygenic scores to identify future pupil performance

To investigate the potential performance of polygenic scores for correctly identifying individual high achieving 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 achieving pupil will be correctly identified as a high achiever) 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 achievers) of a classifier 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 achievers 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 accurately identify most of the pupils who will go on to be in the top 10% of achievers, one would have to set the classification at the point where almost all students would be identified. These results were consistent when other cut-offs were used to determine the high achieving group (Supplementary Figure S3), suggesting that the results do not reflect our definition of high achievers.

Figure 5: ROC curve for being a high achieving student (defined as the top 10% of pupils) at age 7. High achievers defined as pupils with age 16 educational exam scores in the top 10% of the sample. Parental educational attainment (EA) was measured as average years of completed education. Parental socioeconomic position (SEP) was measured as highest parental score on the Cambridge Social Stratification Score scale. Polygenic scores (PGS) built using only genome-wide significant SNPs (GWAS sig SNPs) or all education associated SNPs (All SNP PGS) from the largest GWAS of educational attainment[1]. All PGS analyses include adjustment for the first 20 principal components of population stratification. Note that x axis displays 1-specificty.

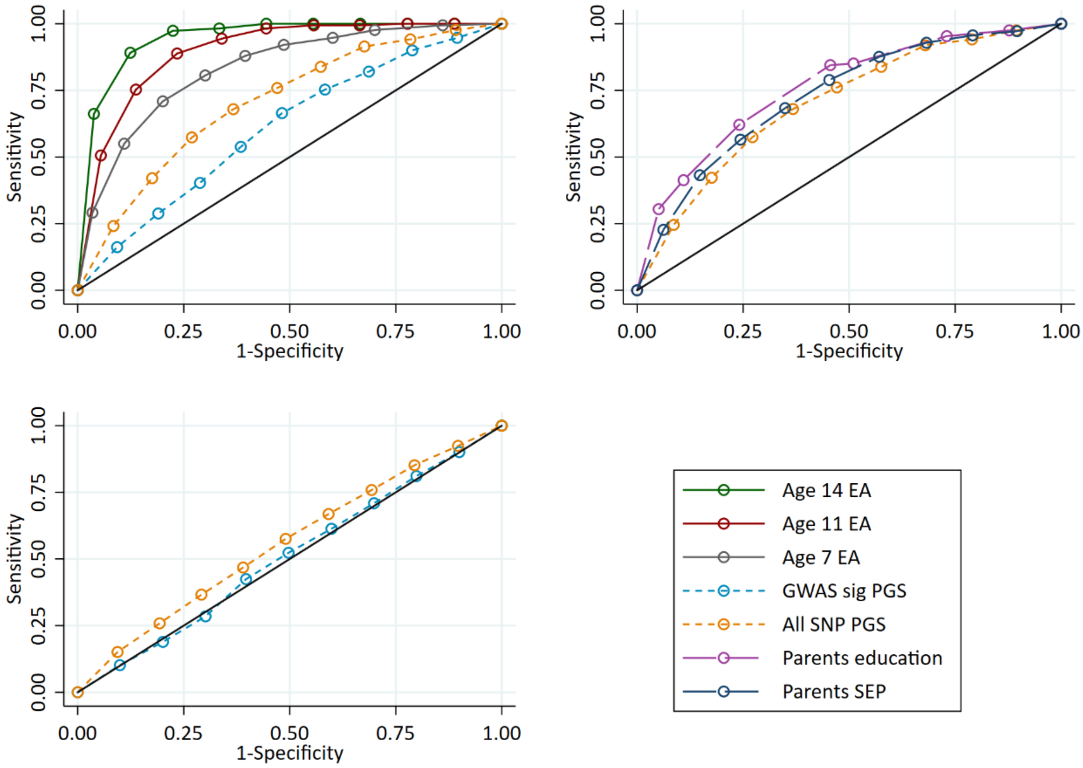


For educational achievement at age 16 when prior achievement data are available, Figure 6a displays that these measures provide far higher sensitivity and specificity than the polygenic scores (AUCs: 0.83 to 0.95 for prior achievement compared to 0.61 to 0.70 for the polygenic scores). That is, there is a far better trade-off between sensitivity and specificity for prior achievement at age 14 than for either polygenic score. For example, a classification point can be set for prior achievement 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 achievement 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). To investigate the use of polygenic scores above phenotypic data, we calculated ROC curves for the polygenic scores on educational achievement at age 16 residualised on age, sex, prior achievement, and pupil characteristics to test the added value of polygenic scores above other phenotypic data. The results (Figure 6c) demonstrated that after accounting for the phenotypic information already available to schools, the polygenic scores do not reliably identify high achievers (AUC: 0.51 and 0.56 for the polygenic scores). The results were consistent had these predictions been made earlier in schooling where later measures of prior attainment were unavailable (Supplementary Figure S4;

AUC's: 0.54 to 0.61). As with achievement at age 7, these results were consistent when other cut-offs were used to determine the high achieving group (Supplementary Figure S5).

If a school headteacher or principal wanted to use polygenic scores as a selection criterion to select the highest performing students, would they identify a group that has higher educational attainment at age 16 than when that selection had been made on other criteria? If they selected the students with the top 10% of polygenic scores, they would on average only sample 24% of the top 10% highest achievers at age 16, and 76% of those not in the top 10%. In contrast, if the principal or policy maker used phenotypic measures from age 11, they would sample 51% of the top 10% highest achievers at age 16, and 49% of those not in the top 10%. This suggests that polygenic scores cannot be used to identify high achieving students more accurately than available phenotypic measures. The group of pupils with the highest polygenic scores do - on average - have higher achievement, but the predictive information provided from the polygenic scores is inferior to that provided by phenotypic predictors. Supplementary Figure S2 demonstrates the variability in age 16 educational achievement for pupils predicted to be in the top 10% from each of the genotypic and phenotypic predictors.

Figure 6: ROC curves for being a high achievement student (pupils with age 16 educational exam scores in the top 10% of the sample) at age 16. Panel A: Independent ROC curves for deciled measures of prior achievement and polygenic scores (PGS) predicting high educational achievement (EA) at age 16. Panel B: Independent ROC curves for deciled measures of parental education and socioeconomic position predicting high educational achievement (EA) at age 16. Panel C: ROC curves for deciled polygenic scores predicting high educational achievement (EA) at age 16 residualised on age, sex, prior achievement, and pupil characteristics available to schools. Parental educational attainment was measured as average years of completed education. Parental socioeconomic position (SEP) was measured as highest parental score on the Cambridge Social Stratification Score scale. Polygenic scores (PGS) built using only genome-wide significant SNPs (GWAS sig SNPs) or all education associated SNPs (All SNP PGS) from the largest GWAS of educational attainment[1]. 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 achievement 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 achievement, accounting for 3.4% and 12.9% of variance (above age and sex) across our sample at age 7 and 16 respectively. The age 16 prediction is higher than the 9.2% reported for high school GPA in the original GWAS[1]. For informative education predictions at the individual level, the most predictive measure was prior achievement. This reflects some current schooling practices whereby pupils are streamed into different classes based upon ability. Conditional on prior achievement there was little incremental gain in the predictive power of polygenic scores for subsequent achievement, suggesting that when prior achievement data are available, polygenic scores are of little utility to providing accurate predictions of a child's future achievement. When children start school and prior achievement data are unavailable, or in cases where pre-intervention characteristics are limited[28], the scores may provide a small amount of predictive power. However, parental socioeconomic position and education were more strongly predictive of achievement than a pupil's genome. Genetic data from individuals therefore provided little information on their future achievement over phenotypic data that is either available or easily obtainable by educators. This is consistent with results from the only other study we are aware of to assess incremental variance explained over parental social characteristics, which observed higher variance explained in years of education by parental education (18% to 21.3%) than the polygenic score (10.6% to 12.7%) in two US samples[1].

The lack of genotypic predictive power that we observed over and above phenotypic data may be because prior achievement mediates the effects of the genotypes on educational outcomes; genetic variants that affect educational achievement at earlier ages are likely to also affect achievement at later ages. More powerful polygenic scores may allow for better prediction of educational achievement in the future. It has been suggested that for complex phenotypes, accurate prediction at the individual level may require a polygenic score that explains up to 75% of the total genetic variance of the phenotype[22]. It is therefore possible that polygenic scores for education will require greater explanatory power for accurate individual prediction. However, our polygenic scores were constructed using results from a GWAS of over a million people, meaning that far larger samples will be required. While future studies may lead to polygenic scores that explain a greater amount of variation in education, these may still not provide useful returns to personalised interventions. High incremental variance explained is a necessary pre-requisite for successful intervention, but it is not a guarantee that an actionable intervention will have a large effect. Furthermore, to provide actionable evidence for personalisation at a given age, polygenic scores need to explain variation in educational outcomes over and above available phenotypic at that age. If most or all the educational differences associated with the polygenic score are phenotypically expressed at a given age, then the score is unlikely to be useful for personalisation.

At the individual level, polygenic scores and parental social background provided similar, but relatively imprecise predictions of achievement within our sample. This reflects a wider issue of the different challenges in analysing group and individual level differences[29]: 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 achievers and all others; while pupils with a high polygenic score are more likely to be high achievers, genetics did not determine high achievement. High academic achievement is due to both environmental and genetic factors, including social background[26], teacher bias[30,31], the home and school environment[32,33], and luck[29]. 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, nurture and chance that is important[34,35].

In fields such as medicine, where genetic risk can be of clinical significance for some diseases[36], personalisation based on genotype may offer actionable intervention at the individual level. However, our results demonstrate that even for the purpose of identifying groups of pupils who will be high achievers, polygenic scores offer limited prediction value above phenotypic data in education. The usefulness of genetic data for educational research however lies in investigating group level differences. This has been previously demonstrated for example in assessing the effectiveness of teachers and schools[17,31]; selection differences between schools[37,38]; social mobility over time and space[39], and, in a different context, for performing Mendelian randomization studies of the effects of education on various outcomes[40,41]. Our results demonstrate that while polygenic scores are useful for investigating group differences such as these, they do not provide suitable value for routine use by teachers and schools to predict a pupil's future achievement. There is a wide range of non-genetic information available to teachers as part of their day to day interactions with pupils that are used to inform and personalise 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. Finally, genetic studies are focused heavily on samples of European ancestry[42]. Polygenic scores built from these studies do not perform well when applied to other ancestry groups[43], meaning that their system-wide application to all pupils in an education system could lead to systematic prediction errors and inequalities in schooling.

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 achievement 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)[44]. Future research could investigate this possibility by conducting a GWAS on detailed standardized exam scores on a large sample. Furthermore, it is possible that polygenic scores from a GWAS conducted on change in test scores throughout education may provide higher prediction accuracy over and beyond phenotypic data if there are genetic factors associated with differences in educational progress. 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%[45], which limits the predictive power of genetic measures for education over some other phenotypes[46]. Finally, issues from confounding biases caused by population level phenomena such as population stratification, assortative mating and dynastic effects (genetic nurture)[18,20,44,47] 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 achievement in between individual analyses. Family data are required to further investigate the impact of these baises[48].

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 achievement 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 achievement and 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 [49,50]. 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[49]. 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 S6).

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 achievement

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 achievement 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[51] 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 (n=763,468), and as such are not perfectly comparable to those reported in the published 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 pupil characteristics that are available to schools 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). For dual parent families we use the average of the two parents' years of education, while for single parent families we use the mother's years of education. Finally, we use a continuous measure of socioeconomic position (SEP), the Cambridge Social Stratification Score (CAMSIS). For dual parent families we used the highest of either parents score, while for single parent families we use the mother's score. Parental years of education and CAMSIS were measured when the study participants were in utero.

Statistical analysis

To examine the predictive ability of polygenic scores for educational achievement we ran a series of regression analyses of the polygenic scores on achievement 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 achievement above age and sex; pupil characteristics; and prior achievement. Bootstrapping with 1000 replications was used to estimate confidence intervals for R^2 values. 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 achievers 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. Department for Education and Skills. *A National Conversation about Personalised Learning*. (2004).

9. Hartley, D. Personalisation: The emerging 'revised' code of education? *Oxford Rev. Educ.* (2007). doi:10.1080/03054980701476311

10. Gilbert, C. *et al.* 2020 Vision: Report of the teaching and learning in 2020 Review Group. *HMSO* (2006).

11. for Children, D. & Schools. *The children's plan: building brighter futures*. **7280**, (The Stationery Office, 2007).

12. Maguire, M., Ball, S. J. & Braun, A. What ever happened to...? 'Personalised learning' as a case of policy dissipation. *J. Educ. Policy* (2013). doi:10.1080/02680939.2012.724714

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

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

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

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

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

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

19. Davies, N. M. *et al.* Within family Mendelian randomization studies. *Hum. Mol. Genet.* (2019). doi:10.1093/hmg/ddz204

20. Morris, T. T., Davies, N. M., Hemani, G. & Smith, G. D. Why are education, socioeconomic

position and intelligence genetically correlated? *bioRxiv* (2019). doi:10.1101/630426

21. 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

22. 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

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

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

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

26. 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).

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

28. Rietveld, C. A. *et al.* GWAS of 126,559 individuals identifies genetic variants associated with educational attainment. *Science (80-.).* **340**, 1467–1471 (2013).

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

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

31. 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).

32. 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

33. 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).

34. 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

35. 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

36. Lu, Y. F., Goldstein, D. B., Angrist, M. & Cavalleri, G. Personalized medicine and human genetic diversity. *Cold Spring Harb. Perspect. Med.* (2014). doi:10.1101/cshperspect.a008581

37. 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

38. 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

39. 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

40. Tillmann, T. *et al.* Education and coronary heart disease: Mendelian randomisation study. *BMJ* (2017). doi:10.1136/bmj.j3542

41. Sanderson, E., Davey Smith, G., Bowden, J. & Munafò, M. R. Mendelian randomisation analysis of the effect of educational attainment and cognitive ability on smoking behaviour. *Nat. Commun.* (2019). doi:10.1038/s41467-019-10679-y

42. Mills, M. C. & Rahal, C. A scientometric review of genome-wide association studies. *Communications Biology* (2019). doi:10.1038/s42003-018-0261-x

43. Duncan, L. *et al.* Analysis of polygenic risk score usage and performance in diverse human populations. *Nat. Commun.* (2019). doi:10.1038/s41467-019-11112-0

44. 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

45. 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).

46. 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

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

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

49. 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).

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

51. 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.