ARTICLE

OPEN

Check for updates

Avoiding dynastic, assortative mating, and population stratification biases in Mendelian randomization through within-family analyses

Ben Brumpton ⓘ et al.[#]

Estimates from Mendelian randomization studies of unrelated individuals can be biased due to uncontrolled confounding from familial effects. Here we describe methods for within-family Mendelian randomization analyses and use simulation studies to show that family-based analyses can reduce such biases. We illustrate empirically how familial effects can affect estimates using data from 61,008 siblings from the Nord-Trøndelag Health Study and UK Biobank and replicated our findings using 222,368 siblings from 23andMe. Both Mendelian randomization estimates using unrelated individuals and within family methods reproduced established effects of lower BMI reducing risk of diabetes and high blood pressure. However, while Mendelian randomization estimates from samples of unrelated individuals suggested that taller height and lower BMI increase educational attainment, these effects were strongly attenuated in within-family Mendelian randomization analyses. Our findings indicate the necessity of controlling for population structure and familial effects in Mendelian randomization studies.

[#]A list of authors and their affiliations appears at the end of the paper.

Mendelian randomization is an approach that uses genetic variants as instrumental variables to estimate the causal effects of one trait (the 'exposure') on another (the 'outcome')[1–5]. It has gained popularity due to the recent expansion in the scale of genome-wide association studies (GWAS) and because it can ameliorate bias due to processes of residual confounding and reverse causation that affect most other observational approaches. In order for Mendelian randomization estimates to be valid, the genetic instrument must meet three assumptions: (1) relevance, it must associate with the exposure, (2) independence, there must be nothing that causes both the instrument and the outcome and (3) exclusion, the association of the instrument and the outcome must be entirely mediated via the exposure. Attention has been focused on developing methods to overcome bias in Mendelian randomization studies due to horizontal pleiotropy[6–11], which would violate the exclusion assumption. However, in this paper we focus on the second assumption: independence. We demonstrate how population and familial effects can violate the second assumption, and that traditional family-based methods are well placed to rectify this problem.

Mendel's laws of genetic inheritance provide a rationale for why much genetic variation for a given trait will be independent of the environment and genetic variation for other traits[1,12]. However, environmental and social factors such as assortative mating, dynastic effects, and population structure may affect the distribution of genetic variants for specific traits within populations (see Supplementary Note 1)[13–16]. Figure 1 illustrates the impact of these processes in the context of Mendelian randomization. The commonality amongst all three processes is that they induce a spurious association between the instrumenting variant and the outcome through confounding. Assortative mating occurs when partners are selected on the basis of phenotype[6,17]. For example, couples tend to have more similar education and body mass index than would be expected by chance[18,19]. If assortative mating arises due to individuals with a particular genetic predisposition selecting mates who have a particular genetically influenced phenotype, this can induce spurious genetic associations which can result in biased estimates from Mendelian randomization studies[6]. In addition, social homogamy may lead to people selecting partners who are similar to themselves[20], and this can compound across generations[6]. Dynastic effects can occur when the expression of parental genotype in the parental phenotype directly affects the offspring phenotype. For example, higher educated parents might support their children's education by providing a stimulating environment, being able to afford tutoring for their child, buying homes in better school districts, or paying for private schools. Other relationships including siblings, grandparents, uncles/aunts and cousins which may affect the offspring's phenotype can also be thought of as a likely generally weaker form of dynastic effects. Finally, residual population structure occurs when there are geographic or regional differences in allele frequency relating to a trait of interest that cannot necessarily be controlled for via principal components[13]. Confounding by population stratification[1], in which ancestry is correlated with both phenotypes and genotypes, was a major concern during the early development of Mendelian randomization[1]. However, this fear was gradually assuaged by a decade of GWAS results that were apparently reliable in the face of population structure[21]. GWAS are now performed on a huge scale; as a consequence the problem of population stratification is again of potential concern because the high statistical power of large studies renders them susceptible to bias from very subtle population structure[13,22].

Confounding in genetic association estimates, as induced by population stratification, dynastic effects and assortative mating,

can and has been resolved by using family-based study designs[6,23,24]. For example, in sibling pair studies, genetic associations at loci can be partitioned into between pair and within pair components[23]. Because genetic differences within sibling pairs reflect random independent meiotic events, within pair effects are unrelated to population stratification and most potential confounders that might influence the phenotype. Similarly, other family-based designs and within-family tests to adjust for or exploit parental genotypes exist, such as estimating maternal and offspring genetic effects using structural equation modelling[25], quantitative transmission/disequilibrium tests[26,27], or mother-father-offspring trios to adjust for parental genotypes[28]. Such within-family designs have been used to validate results from GWAS[29,30], obtain unbiased heritability estimates[31], and assess causation in the classical twin design[32,33]. Yet, despite the initial extended proposal of Mendelian randomization advising that the only way to ensure true randomization was through a within-family design[1], to-date contemporary implementations using modern genomic methods have rarely been performed. The principal reason for this has been a lack of genomic data collected from families at a scale sufficient to be suitably powered. As we now enter the age of national scale biobanks and very large twin studies, this essential extension of Mendelian randomization is becoming feasible.

This paper presents theory and simulations that demonstrate how within-family designs can be coupled with genomic data to perform Mendelian randomization analyses unbiased by population structure and family effects. We integrate these approaches in a modular fashion alongside other methods that have been developed for pleiotropy-robust inference (i.e. to be resilient to violations of the third assumption of Mendelian randomization)[7–9,34]. Using 28,777 siblings from HUNT, 32,231 siblings from the UK Biobank, and 222,368 siblings from 23andme we illustrate these methods empirically. First, we estimate the causal effect of BMI on high blood pressure and risk of diabetes as positive controls. Second, we estimate the casual effect of height and BMI on educational attainment and find substantial differences between estimates from unrelated individuals and estimates using within-family-based approaches, demonstrating the importance of controlling for family effects and population structure in Mendelian randomization studies.

Results

Bias evaluated using directed acyclic graphs. There are three mechanisms depicted in Fig. 1 that induce bias in the SNP-outcome relationship of a Mendelian randomization design. The problem of population structure is well known and has been examined in detail in Lawson et al. 2019[35]. A similar confounding structure can be induced through dynastic effects because there is a path between the offspring instrument value and the offspring outcome value, which arises through parental inheritance.

Similarly, cross trait assortative mating can induce bias due to a form of collider bias, where conditioning on the assortment of parents induces a correlation between the SNP effects on x and all other genetic effects on y. As this has been demonstrated in simulations before[6], below we demonstrate the utility of within-family designs for protecting MR estimates from bias due to familial effects, and then illustrate their importance using empirical examples. We report the analytic bias terms for the dynastic effects and assortative mating in Supplementary Notes 2 and 3.

Simulations demonstrates robustness of the within-family design. We conducted forward-in-time simulations to investigate bias and power related to estimating the causal effect of an
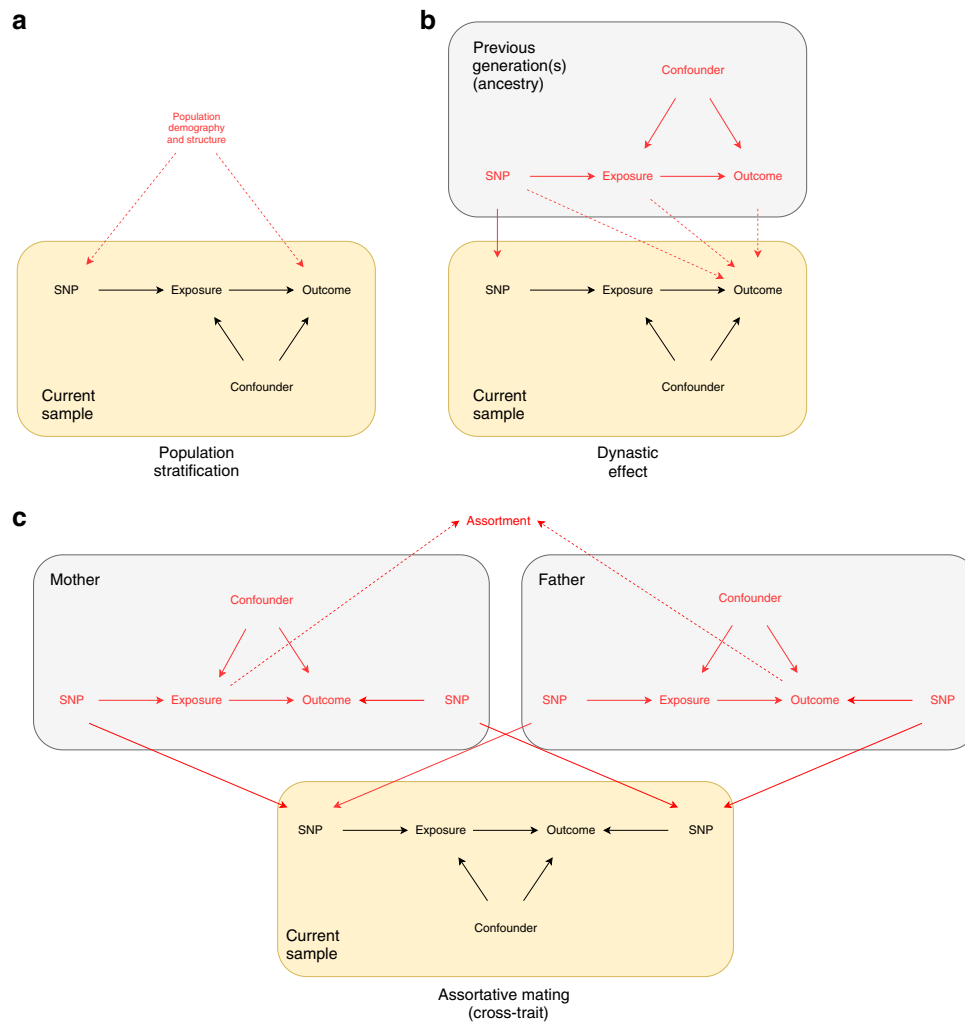
Fig. 1 Directed acyclic graphs illustrating how population structure and familial effects can cause confound MR studies. Black arrows indicate causal paths in the index individual, red arrows indicate causal paths in the parents, and dashed red arrows indicate confounding paths. The MR estimate of the causal effect of the exposure on the outcome is biased because of potentially unobserved confounders between the SNPs and the exposure and the outcome. **a** Illustrates how population demography and structure can confound the SNP-outcome association. **b** Illustrates how dynastic effects can induce the same statistical confounding structure of the SNP-outcome association through an entirely different mechanism. The solid red vertical arrow indicates the genetic inheritance of germline DNA. The dotted line indicates the direct (dynastic) effect of the parents on the offspring's outcomes. These can either be mediated via the exposure, the outcome or some other mechanism indicated by the direct arrow from SNP to offspring outcome. MR estimates of the effect of the exposure on the outcome in samples of unrelated individuals will be biased because there is a path between offspring SNP and the outcome via the effect of the parents' phenotypes on their offspring's outcomes (dynastic effects). The presence of dynastic effects would violate one of three key MR (instrumental variable) assumptions—the independence assumption. Estimates that control for mother or father genotype, or sibling genotype will close this path and be unbiased. **c** Illustrates how assortative mating is a third mechanism that can confound the SNP-outcome association. In this example we present cross-trait assortative mating where there is a pathway between the mother's genotype and offspring's outcome via the father's genotype for the outcome. All these forms of SNP-trait confounding can be accounted for by using methods based on within-family contrasts.

exposure on an outcome. In the simulations each parent transmits genotypes to their offspring, and the parents' exposure had a direct causal effect on the offspring's outcome (Fig. 1b). In null simulations where the exposure effect on the outcome was zero, the Mendelian randomization estimates using unrelated individuals were biased and had high false discovery rates in the presence of dynastic effects (false discovery rate > 0.75 when the confounders were C_x and $C_y = 0.1$, $b_{ux} = 0.1$, $n > 10,000$). The pattern of bias in the sibling and trio methods was substantially improved, with a small amount of weak instrument bias observed, which attenuated as sample sizes improved (Fig. 2).

Where we simulated the exposure to have a causal effect on the outcome Mendelian randomization using unrelated individuals had the highest power (Fig. 2). However, the sibling

and trio design also performed well with larger sample and effect size (power > 0.9 when sample sizes $\geq 10,000$, dynastic effect ≤ 0.2, effect size $= 0.05$). The within-family models were substantially less powerful than Mendelian randomization using unrelated individuals; as usual, controlling bias comes at a cost.

Empirical study using HUNT, UK Biobank and 23andMe. We begin our empirical study by using two positive control MR analyses to estimate the causal effects of two well established causal effects: BMI on diabetes and BMI on high blood pressure. We ran these analyses with and without allowing for a within family effect.
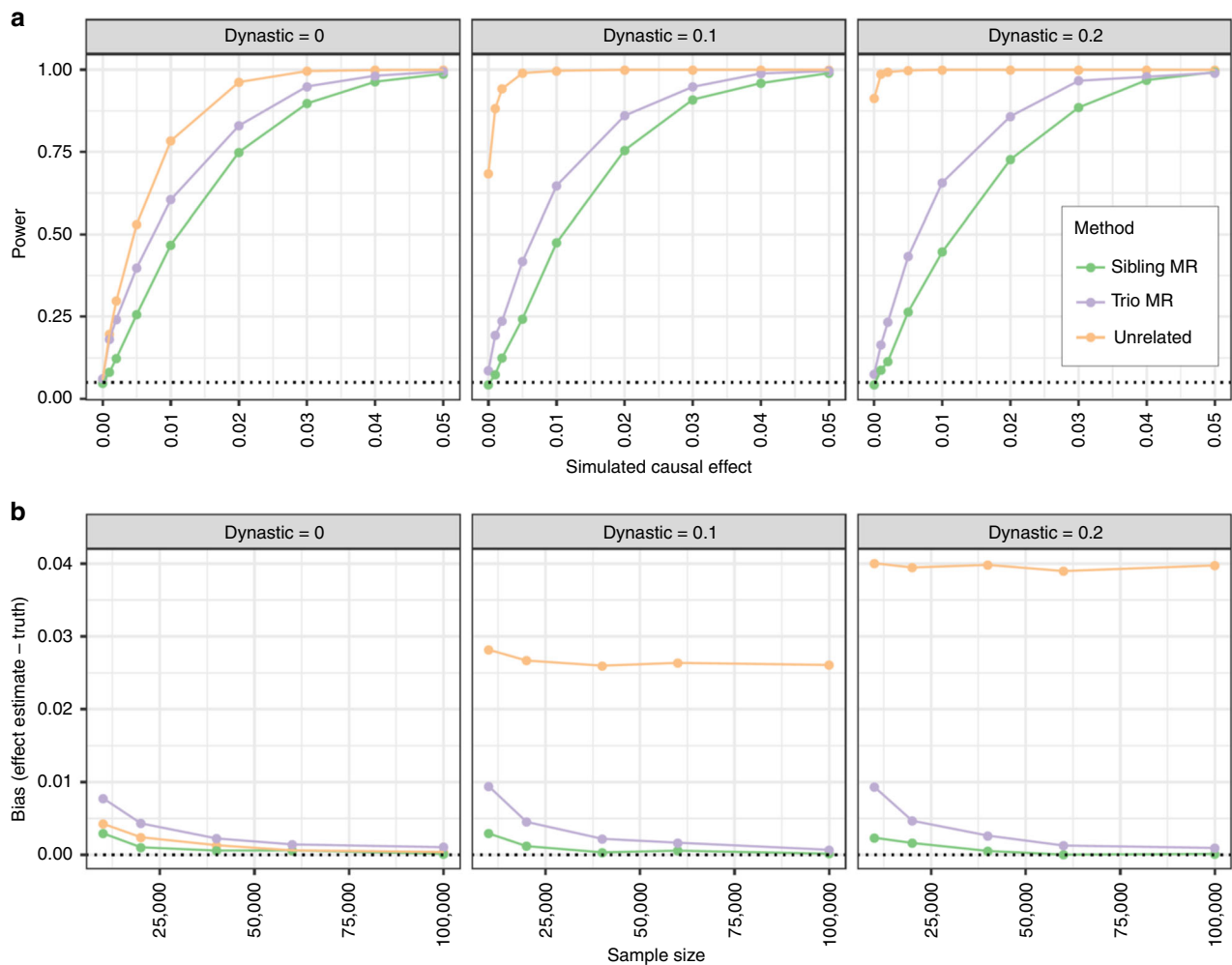
Fig. 2 Results of simulations comparing different Mendelian randomization study designs for power and bias. a SNP-exposure $r^2 = 0.05$; sample size = 10000 singletons, sibs, or trios; simulation involves an influence of parental exposure influencing child's confounder, which explains 10% of variance in child exposures and outcomes. For a simulated causal effect = 0, we expect the false discovery rate to be 0.05. **b** Estimated bias by sample size using different Mendelian randomization designs. The simulations are similar to **a** but allow sample size to vary and fixing the causal effect of an exposure x on an outcome y to 1% of variance explained. The bias in within-family Mendelian randomization estimates is slightly elevated when sample sizes are small due to weak instrument bias, but are otherwise are protected from the large bias seen when using unrelated samples.

Participants with higher BMI were more likely to have diabetes: each $1 \, \text{kg/m}^2$ higher BMI was associated with a 0.60 (95%CI: 0.55–0.65, p-value $< 1.2 \times 10^{-136}$) percentage point increase in the diabetes risk. These differences were modestly attenuated after including a family fixed effect (0.46, 95%CI: 0.40–0.52, p-value = 8.5×10^{-52}). The Mendelian randomization estimate using unrelated individuals suggested that each unit increase in BMI increased the risk of having diabetes by 0.82 (95%CI: 0.71–0.93, p-value = 3.3×10^{-50}) percentage points. This estimate remained after allowing for the fixed effects of family (1.01 percentage point increase per $1 \, \text{kg/m}^2$ increase in BMI, 95%CI: 0.58–1.44, p-value = 3.3×10^{-06}). The summary data Mendelian randomization analysis allowing for family effects estimates were similar (0.75 percentage point increase per $1 \, \text{kg/m}^2$ increase in BMI, 95%CI: 0.38–1.13, p-value = 7.6×10^{-05}, $p_{\text{diff unrelated}} = 0.74$). On average the associations of the SNPs with BMI and diabetes were similar before and after allowing for a family fixed effect, falling 7% (95% CI: −5% to 20%, p-value = 0.26) and increasing 11% (95%CI: −17% to 40%, p-value = 0.42) respectively.

Participants with higher BMI were more likely to have high blood pressure; each $1 \, \text{kg/m}^2$ higher BMI was associated with a 2.63 (95%CI: 2.54–2.72, p-value $< 1 \times 10^{-300}$) percentage point increase in high blood pressure risk. This association did not attenuate after including a family fixed effect (2.42, 95%CI: 2.30–2.54, p-value $< 1 \times 10^{-300}$). The Mendelian randomization estimate using the sample of unrelated individuals suggested that each unit increase in BMI increased the risk of having high blood pressure by 1.59 (95%CI: 1.34–1.83, p-value = 1.3×10^{-36}) percentage points. The Mendelian randomization estimate was similar after allowing for a family fixed effect (1.13 percentage point increase per $1 \, \text{kg/m}^2$ increase in BMI, 95%CI: 0.04–2.21, p-value = 0.04). The summary data Mendelian randomization estimates were similar (0.76 percentage point increase per $1 \, \text{kg/m}^2$ increase in BMI, 95%CI: −0.19 to 1.70, p-value = 0.12, $p_{\text{diff unrelated}} = 0.10$). On average the associations of the SNPs and high blood pressure fell by 51% (95%CI: 23–80%, p-value = 0.0006) after allowing for family fixed effects. Overall, the associations found from the within-family studies for these positive controls establish the utility of the study design with the current scale of data.

Next, in order to examine the contrast between MR methods using unrelated individuals and family-based designs, we investigate two associations that are more liable to bias due to population structure and familial effects: height on years of
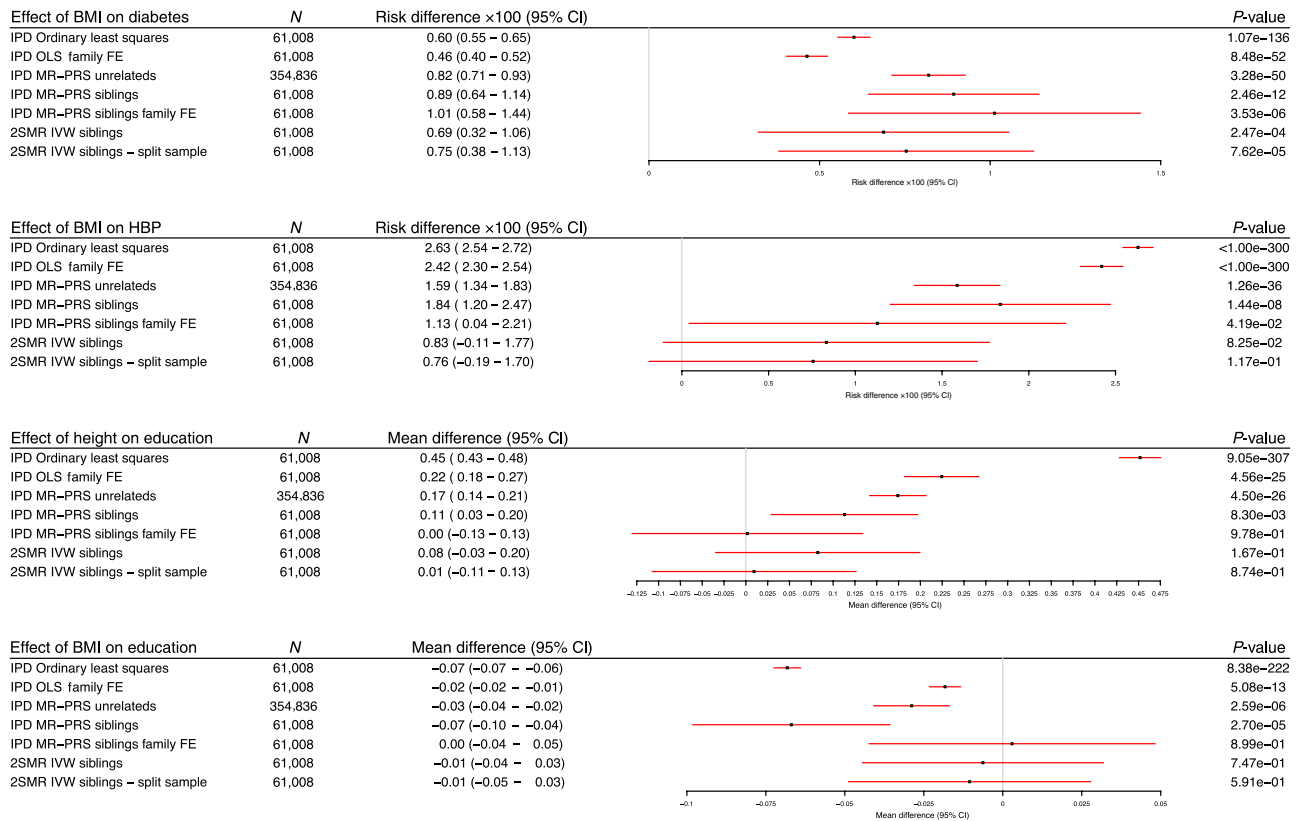
Fig. 3 Estimates of the effect of BMI on self-reported diabetes and high blood pressure and height and BMI on educational attainment using ordinary least squares, Mendelian randomization in unrelated individuals and samples of siblings, point estimates and 95% confidence intervals reported. All methods were consistent with higher BMI increasing diabetes and high blood pressure risk. Being taller and having lower BMI were observationally associated with higher educational attainment. The effects of height and BMI on educational attainment were attenuated but still apparent when using Mendelian randomization estimates based on unrelated individuals from HUNT and UK Biobank. The effects were eliminated after allowing for a family effect using individual-level or summary data Mendelian randomization.

education, and BMI on years of education. Taller participants were more educated; each 10 cm increase in height was associated with an additional 0.45 (95%CI: 0.43–0.48, p-value $= 9.0 \times 10^{-307}$) years of education (Fig. 3). This association was attenuated after including a family fixed effect (0.22, 95%CI: 0.18–0.26, p-value $= 4.6 \times 10^{-25}$). The Mendelian randomization estimate using the sample of unrelated individuals implied that each 10 cm increase in height caused an increase of 0.17 (95%CI: 0.14–0.20, p-value $= 8.5 \times 10^{-26}$) years of education. After allowing for a family fixed effect, the Mendelian randomization estimate was greatly attenuated suggesting little evidence of a causal effect of height on education (mean difference per 10 cm increase in height: 0.002, 95%CI: −0.13 to 0.13, p-value $= 0.98$). When we used two sample Mendelian randomization by estimating the SNP-exposure and SNP-outcome associations in different samples (split sample)[36,37] and then meta-analysing, there was little evidence of a causal effect of height on education (mean difference per 10 cm increase in height $= 0.009$, 95%CI: −0.11 to 0.13, p-value $= 0.87$, $p_{\text{diff unrelated}} = 0.008$). On average, the associations of these SNPs with height and education fell by 18% (95%CI: 14–22%, p-value $= 8.5 \times 10^{-24}$) and 61% (95%CI: 49–73%, p-value $= 1.5 \times 10^{-21}$) after allowing for family fixed effects, respectively.

On average, participants with higher BMI were less educated: each 1 kg/m² higher BMI was associated with 0.07 fewer years of education (95%CI: 0.06 to 0.07, p-value $= 8.4 \times 10^{-222}$, see Fig. 3). This association was attenuated after including a family fixed effect (0.02, 95%CI: 0.01 to 0.02, p-value $= 5.1 \times 10^{-13}$). The Mendelian randomization estimate without allowing for

familial effects implied that each additional unit of BMI decreased years of schooling by 0.03 (95%CI: 0.02–0.04, p-value $= 2.6 \times 10^{-06}$). This effect was eliminated after allowing for a family fixed effect, providing little evidence for a causal effect of BMI on educational attainment (mean difference per 1 kg/m² higher BMI $= 0.00$, 95%CI: −0.04 to 0.05, p-value $= 0.89$). Again, the effect was also largely attenuated when we used two sample summary data approaches. Using separate samples to estimate the SNP-exposure and the SNP-outcome associations allowing for family fixed effects, there was little evidence of an effect of BMI on educational attainment (mean difference per 1 kg/m² higher BMI $= -0.01$, 95%CI: −0.05 to 0.03, p-value $= 0.59$, $p_{\text{diff unrelated}} = 0.002$). On average, the association of the 69 BMI SNPs and education fell by 65% (95% CI: 34–76%, p-value $= 1.8 \times 10^{-06}$) after allowing for family fixed effects. These results suggest that the methods that do not account for familial effects may be biased estimators of the individual level causal effect. We found little evidence of heterogeneity between the two sample Mendelian randomization estimates from UK biobank and HUNT, except for the effect of BMI on diabetes (p-value $= 0.027$).

We investigated whether our results could be explained by pleiotropy using the weighted median, weighted modal and MR-Egger estimators. These summary data Mendelian randomization estimators use estimates of the SNP-exposure and SNP-outcome associations to estimate the effect of the exposure on the outcome. These estimators are robust to a number of forms of pleiotropy. There was little evidence of differences between the inverse variance weighted (IVW) and pleiotropy robust methods,

pleiotropy from the MR-Egger intercept, or heterogeneity across the studies (Supplementary Fig. 1).

We investigated the difference (shrinkage) of the total to within family SNP-phenotype associations in HUNT and UK Biobank using seemingly unrelated regression (SUR). The estimated shrinkage is given in Supplementary Table 1. The shrinkage of the estimates suggests that accounting for familial effects affects the estimated SNP-phenotype associations for all phenotypes analysed. Educational attainment was the most strongly affected, falling by 56.8% (95%CI: 49.2–64.4%). Diabetes was the least affected falling by 11.2% (95%CI: 1.3–21.1%).

We replicated our findings using data from 223,368 individuals (111,684 families) sampled by 23andMe (Supplementary Fig. 2). There was evidence that a 1-kg m^{-2} higher BMI increased the absolute individual level liability to diabetes and high blood pressure by 0.69 (95%CI: 0.37–1.00) and 1.26 (95%CI: 0.90–1.63) per 100 people, respectively. There was little evidence of heterogeneity across the weighted median, modal or MR-Egger estimators. This suggests that under a range of assumptions about pleiotropy and controlling for familial effects, BMI is likely to increase risk of diabetes and high blood pressure. In contrast, there was little evidence that height or BMI had a substantial effect on educational attainment. The results are precise and suggest that each 10 cm taller height is unlikely to increase years of education by more than 0.02 years (mean difference = 0.00, 95%CI: −0.01 to 0.02), and a 1 kg/m^2 unit higher BMI is unlikely to decrease years of education by more than 0.02 years (mean difference = 0.00, 95%CI: −0.02 to 0.02). The weighted median, modal and MR-Egger estimators were consistent with the IVW estimates.

Discussion

We have presented within-family methods for Mendelian randomization and demonstrated how confounding due to population structure and familial effects can bias Mendelian randomization studies using unrelated individuals. As with most instrumental variable estimators, ceteris paribus the size of the bias induced by familial effects will be larger the smaller the individual level causal effect of the genetic variant on the exposure. The simulations illustrated how bias occurs even if the phenotype of interest has no direct causal effect on the outcome, and that these effects can theoretically induce false positive findings. The simulations further demonstrated how samples of related individuals can be used to control for these effects either using siblings or parent-offspring trios. These designs can be used in conjunction with existing approaches for accounting for horizontal pleiotropy, another potential source of bias to arise in Mendelian randomization studies. However, estimates from within-family Mendelian randomization are less precise than estimates using unrelated individuals, which is consistent with those seen for allelic association[38–40]. Compounding the issue of statistical power, there are fewer relatives than unrelated individuals in most studies. In samples from HUNT, UK Biobank and 23andMe, we investigated the impact of population structure and familial effects on four empirical examples; the effects of BMI on the risk of diabetes and high blood pressure and the effects of height and BMI on educational attainment. We found that the effects of BMI on the risk of diabetes and high blood pressure were less precise, but consistent when allowing for family effects. Conversely, the effects of height and BMI on educational attainment were almost entirely attenuated after allowing for family fixed effects, suggesting that results from previous Mendelian randomization studies using unrelated individuals may have been biased.

A substantial literature has used Mendelian randomization and samples of unrelated individuals to establish that BMI increases the risk of diabetes and high blood pressure later in life[41]. Our results suggest that confounding due to familial effects is unlikely to explain these results, and that they are more likely due to an individual level causal effect of BMI on an individual's risk. Behavioural geneticists have used longitudinal data from samples of twins to understand how different family members affect each other over time[42,43]. Other studies have used animal models to investigate how social genetic effects (i.e. indirect or dynastic effects) can affect health outcomes[44]. A rich literature has established that height and BMI are respectively positively and negatively associated with educational attainment and socio-economic position[45–47]. Consistent with our results, previous studies using twin data have indicated that the relationship between height and educational attainment is likely to be due to familial effects[48,49]. These findings raise questions about whether height and BMI have individual level causal effects on socio-economic outcomes later in life[50–52]. Our results indicate that familial effects can have important phenotypic consequences on widely studied relationships such as between height and BMI and education.

In general, within-family Mendelian randomization estimates are less precise than estimates from samples of unrelated individuals. Thus, within-family estimates of a specific association can be considered more robust, but less efficient estimates. Therefore, if there is evidence of differences between the estimates, then generally the more imprecise but less biased within-family estimates should be preferred. Our estimates of the effect of height and BMI on educational attainment are an example of this situation. If there is little evidence of differences between estimates using unrelated individuals and those allowing for family effects, then the former estimates should be preferred. Our estimates of the effect of BMI on risk of diabetes and high blood pressure are an example of this situation. This is analogous to comparing instrumental variable estimates to multivariable adjusted estimates[3]. While allowing for family fixed effects or using difference estimators will account for dynastic effects or assortative mating, these methods will not address bias due to violations of the third Mendelian randomization assumption (exclusion restriction). This assumption is that the SNPs have no direct effect of the SNPs on the outcome (i.e. no pleiotropy). MR-Egger, weighted median and mode, or Lasso estimators are robust to various forms of violations of this assumption[7–9,34]. It is trivial to use these estimators with the summary data methods we describe above and illustrate in Supplementary Figure 1. However, typically these estimators have lower power than the IVW estimator. The within-family summary data SNP-exposure and SNP-outcome associations, which allow for a family fixed effect, can be used with existing summary data estimators. Other proposed approaches allow for sophisticated control and estimation of pleiotropy and can trivially include family fixed effects[53], but again generally have lower power and require more data than Mendelian randomization approaches using allele scores or IVW. Supplementary Table 2 contains the ratio of standard errors for the MR analyses using unrelated individuals and allowing for a familial effect for the empirical results for the MR using individual participant data (IPD) and polygenic scores (MR-PRS) and two sample Mendelian randomization (2SMR) using an IVW. These estimates use identical samples, but the within family estimates had standard errors that were between 23% and 71% larger. This implies that the within family analyses would require total samples sizes between 150% and 294% larger in order to match the power of current sample sizes of unrelated individuals. A further issue concerns residual population stratification in ancestrally heterogenous GWAS such as GIANT, which may bias

SNP-phenotype associations for height, and affect analyses in other samples using SNPs identified in those GWAS[22]. Within-family Mendelian randomization can control for residual population stratification.

Within-family estimates from samples of siblings that allow for a family fixed effect are robust to biases due to dynastic effects, assortative mating and fine population structure[1,54,55]. Unlike analyses using summary data from unrelated individuals, two sample within family designs do not require the familial effects to be the same in the two samples. This is because the (different) familial effects in each sample are controlled for and the MR estimates use the individual level causal effect. Of family-based approaches, the sibling design is potentially most useful because large amounts of such data are available through biobanks and family-based studies. A limitation of current sibling designs is that they assume no sibling-sibling interaction effects. Phenotypic similarity of siblings may reflect 'passive' sharing of environments or genes, or 'active' imitation or contrast effects arising from interaction between siblings[56]. Contrast effects, which may inflate the estimated contribution of the nonshared environment in twin studies[57], can be mimicked by parental rating bias[58,59]. However, for biological phenotypes where rating bias is not a concern, Mendelian randomization could be used to study the influence via imitation or contrast of one sibling's genotype on the other's phenotype, sometimes called 'social genetic effects'[44], thereby adding to work on dynamic interplay between siblings[42,43].

Population structure, dynastic effects and assortative mating may cause bias in GWAS[14]. If a GWAS is aiming to estimate the causal effect of variants on a given phenotype, then samples of unrelated individuals may produce biased estimates and potentially spurious findings. Population structure and dynastic effects can cause bias under the null hypothesis of no effect i.e. induce spurious false positive signals. Single trait assortative mating will, however, be an unbiased test of the null hypothesis that the SNP does not affect the phenotype but will inflate SNP-phenotype associations. However, cross trait assortative mating can cause bias under the null hypothesis. Future studies could re-run GWAS on a full range of traits on samples of siblings allowing for family fixed effects. This approach would also address concerns about residual population stratification in GWAS, which may bias SNP-phenotype associations in GWAS including populations with heterogeneous ancestry[22]. However, to detect genetic variants that explain 0.1% of the variance of either the offspring or maternal effects (i.e. a 2 df test) will require sample sizes of 50,000 mother-offspring pairs to detect GWAS ($\alpha = 5 \times 10^{-08}$). Sample sizes of around 10,000 will be required to partition known loci of similar size to the above into maternal and/or offspring genetic effects ($\alpha = 0.05$)[60]. This sample size would provide valuable information about which phenotypes are likely to be most strongly affected by dynastic effects and assortative mating. It is likely that many, particularly biological, traits are relatively unaffected by these effects and thus GWAS results for these traits are unlikely to be biased due to these factors, how this requires investigation. Recent GWAS of social traits such as education reported the attenuation after allowing for family effects in their estimates in small samples[30]. Further work in this area should include estimating the consequences of familial effects for GWAS and Mendelian randomization estimates.

Population structure and familial effects can cause bias in Mendelian randomization studies. We found differences between estimates from unrelated individuals and within-family estimates in simulations and empirical analysis. The causal estimates of the effect of height and BMI on educational attainment were almost entirely attenuated after allowing for family fixed effects. Within-family methods, either using individual-level, or summary data Mendelian randomization approaches can be used to obtain unbiased estimates of the causal effects of phenotypes in the presence of dynastic effects, assortative mating and population stratification.

Methods

Statistical models. We describe four methods of using family data for Mendelian randomization below. If there are only two siblings, the difference and family fixed effects methods are equivalent, see appendix for proof.

The model to be estimated can be described as:

$$x_{k,i} = \gamma_0 + \gamma_1 g_{k,i} + \gamma_2 C_{k,i} + f_k + v_{k,i} \tag{1}$$

$$y_{k,i} = \beta_0 + \beta_1 x_{k,i} + \beta_2 C_{k,i} + f_k + u_{k,i} \tag{2}$$

where $y_{k,i}$ and $x_{k,i}$ are the outcome and exposure for individual i from family k. $g_{k,i}$ is a set of genetic variants that are associated with the exposure. f_k is a family-level confounder, modelled in the empirical analysis via a family fixed effect (i.e. an indicator variable for each family). This accounts for all time invariant family-level confounders of the genetic variant-outcome association. Both $g_{k,i}$ and f_k are functions of a family-genetic component. $u_{k,i}$ and $v_{k,i}$ are random error terms. $C_{k,i}$ is a confounder of the association of the exposure and the outcome, γ_2 and β_2 indicate the effect of the confounder on the exposure and the outcome. β_1 is the true causal effect of the exposure on the outcome which we wish to estimate. This model implies that Mendelian randomization using data from unrelated individuals would produce biased estimates of β_1 due to the correlation between $g_{k,i,j}$ and f_k. The effect of the exposure on the outcome can be estimated using individual level data allowing for a family fixed effect, or summary level data using difference methods within families, or by allowing for a family fixed effect. We describe these approaches below.

Siblings difference method. To apply Mendelian randomization to samples of siblings, effect estimates for the SNP-exposure association and SNP-outcome association are based on correlating the phenotypic divergence with the genotypic divergence within sibling pairs. Taking the difference between siblings removes the effect of the family-level confounder. For any pair of siblings within family k, indicated $k,1$ and $k,2$, the genotypic difference at genetic variant j is:

$$\delta_{k,j} = g_{k,1,j} - g_{k,2,j} \tag{3}$$

The association between the genotypic differences and phenotypic differences in the exposure, x, and outcome y, for SNP j can be estimated via:

$$x_{k,1} - x_{k,2} = \gamma_j \delta_{k,j} + \dot{v}_{k,j} \tag{4}$$

$$y_{k,1} - y_{k,2} = \Gamma_j \delta_{k,j} + \dot{u}_{k,j} \tag{5}$$

The estimated associations, γ_j and Γ_j, can be used with any summary level Mendelian randomization estimator. Here we apply the inverse variance weighted (IVW) approach. Each pair of siblings can be included as a separate pseudo-independent pair.

Family fixed effect with sibling data. Alternatively, we can estimate the associations using family fixed effects indicated by f_k for each family, which is equivalent to centring the data by subtracting the family mean.

$$x_{k,i} = \gamma_0 + \gamma_{1,j} g_{k,i,j} + f_k + \ddot{v}_{k,i,j} \tag{6}$$

and

$$y_{k,i} = \beta_0 + \Gamma_1 g_{k,i,j} + f_k + \ddot{u}_{k,i,j}. \tag{7}$$

This estimator accounts for any differences between families, which includes any effect of assortative mating or dynastic effects common to all siblings by including a dummy variable for each family. This provides unbiased estimates of the SNP-exposure and SNP-outcome associations. These estimates can be used with standard summary data Mendelian randomization methods. The difference and family fixed effects methods are identical if there are only two siblings in each family. This fact follows from substituting equations iv and v into equations ii and iii and simplifying (see "Appendix" for proof). If there are more siblings, then the estimators are non-identical, but likely to be similar, see the appendix for further details. An analytically convenient method to use for this estimator is the within transformation. The within transformation either de-means the variables for or additionally adjusts for the family level means. Demeaned using the within transformation is computationally efficient, particularly for large sample sizes—and is the analytic method used by many statistical packages for fixed effects estimators. An advantage of further adjusting for the within family mean is that it provides an estimate of the between family effect. Cluster robust standard errors can be used to allow for clustering and relatedness within families.

Adjusting for parental genotype with mother-father-offspring trio data. Finally, if data on mother-father-offspring trios are available, the estimates of the SNP-exposure and SNP-outcome associations for each child can be adjusted for

their mother's and father's genotypes, indicated by $g_{im,j}$ and $g_{if,j}$ respectively[61]:

$$x_i = \gamma_0 + \gamma_{1,j}g_{i,j} + \gamma_{2,j}g_{im,j} + \gamma_{3,j}g_{if,j} + u_{i,j} \qquad (8)$$

and

$$y_i = \beta_0 + \Gamma_1 g_{i,j} + \Gamma_2 g_{im,j} + \Gamma_3 g_{if,j} + v_{i,j} \qquad (9)$$

Again, these associations can be used to estimate the effect of the exposure on the outcome using summary data Mendelian randomization methods. It is possible to estimate the effect of offspring genotype on the exposure and outcome conditional on the mother and father genotype using summary data[25,61]. The estimated causal effect can be biased if both the SNP-exposure and SNP-outcome associations are estimated in the same sample[62]. This bias can be eliminated by splitting the sample and estimating the associations in separate samples.

Two-stage least squares with sibling data. Many summary data methods assume no measurement error on the SNP-exposure association (NOME)[63]. This assumption may lead to underestimation of the standard error of the effect of the exposure on the outcome. Two-stage least squares can estimate the effect of the exposure on the outcome using the individual-level data from siblings. Estimators that use individual-level data can integrate the estimation error from the SNP-exposure association. We used cluster robust standard errors that allow for clustering and relatedness within family. We used the commands xtivreg and plm[64].

Simulation of dynastic effects. We simulated a cohort consisting of pairs of unrelated mothers and fathers who had two offspring. All individuals had a genome of 90 SNPs. We set the distribution of identity by descent (IBD) across the 90 SNPs as N (0.5,0.037) for each sibling pair, as per theory, because there are on average 90 recombination events separating human siblings. Hence, we assume that each SNP has an independent effect on the exposure.

We defined parents' exposure and outcome by defining confounders u, exposure x and outcome y. A directed acyclic graph illustrating these relationships is shown in Fig. 1b. The confounder influences the parents' exposure and outcome. The offspring have the same confounding structure, except the parent's exposure affects their offspring's outcomes via a dynastic or ancestry effect. The genetic influence of each of the 90 SNPs on the exposure amounts to explaining V_{gx} of the variance in the exposure. We assumed no horizontal pleiotropy. All estimates assume $V_{gx} = 0.1$ and 90 independent causal variants (i.e. somewhat similar to GIANT results for BMI)[65].

To generate the phenotypes under a model of dynastic effects, the offspring outcome was influenced by both the offspring exposure and the parents' exposures. In these simulations all phenotypes had mean of 0 and variance of 1. Differing strengths of dynastic effects by which the parental exposure influenced the offspring outcome were generated ($b_{ux} = 0,0.01,0.02$) under a set of models with a range of causal effects of the exposure on the outcome ($b_{xy} = 0,0.001,0.002, 0.005,0.01,0.05$). We calculated the false discovery rate (proportion of test with p-value < 0.05) for 100 iterations of each simulation using each of three methods: standard IVW as applied to one of each individual in a set of siblings (i.e. a sample of unrelated individuals), the within-family sibling design, and the within-family trio design. Finally, we calculated bias (estimated effect— simulated effect) for all three study approaches by simulated confounding (C_x and $C_y = 0,0.1,0.2$), dynastic bias ($b_{Ux} = 0,0.1,0.2$) and simulated causal effects ($b_{xy} = 0,0.001,0.002,0.005,0.01,0.05$). The sample sizes were 10,000, 20,000, 40,000, 60,000 and 100,000 sibling pairs for all simulations.

If the familial effect influences the exposure, but does not affect the outcome, then we would not expect bias in the Mendelian randomization analysis. This is because there would be no open path between the SNP and the outcome. This is illustrated in the directed acyclic graph illustrated in Supplementary Fig. 3.

Empirical analysis. To demonstrate the approach and assess potential bias from population structure and familial effects, we conducted within-family Mendelian randomization using two illustrative examples in the HUNT and the UK Biobank[66–68]. We estimated the effects of BMI on high blood pressure and diabetes and the effects of height and BMI on educational attainment. The effects of BMI on diabetes and high blood pressure have been well studied and provide a positive control[41]. These effects on clinical outcomes experienced later in life are unlikely to be due to assortative mating or dynastic effects, because parents are less likely to assort on genetic liability for diabetes or high blood pressure. The genetic liability for these conditions was probably unknown when the couples were formed. Previous longitudinal and Mendelian randomization studies using unrelated individuals have suggested that height and BMI may affect educational attainment[45,50]. Such an association, if causal, might be counteracted by changing educational policy. However, the association may be due to parents' education, via dynastic effects or assortative mating, where more educated people select taller and slimmer partners. Assortative mating and dynastic effects can confound the association between genetic variants when data from the offspring generation are used. Therefore, the ratio of individual-level causal effects to family-level effects is likely to be higher for the effects of BMI on clinical end points than for the effects of height and BMI on education.

The Nord-Trøndelag Health Study. The Nord-Trøndelag Health Study (HUNT) is a population-based cohort study. The study was carried out at four time points over approximately 40 years (HUNT1 [1984-1986], HUNT2 [1995-1997] and HUNT3 [2006-2008] and HUNT4 [2017-2019]). A detailed description of HUNT is available[66]. We include 71,860 participants from HUNT2 and HUNT3 as they have been recently genotyped using one of three different Illumina HumanCore-eExome arrays (HumanCoreExome12 v1.0, HumanCoreExome12 v1.1 and UM HUNT Biobank v1.0). For a flow chart of participants inclusion and exclusion from the study see Supplementary Figure 4. Imputation was performed on samples of recent European ancestry using Minimac3 (v2.0.1, http://genome.sph.umich.edu/wiki/Minimac3) from a merged reference panel constructed from the Haplotype Reference Consortium (HRC) panel (release version 1.1) and a local reference panel based on 2202 whole-genome sequenced HUNT participants[12–14]. Ancestry of all samples was inferred by projecting all genotyped samples into the space of the principal components of the Human Genome Diversity Project (HGDP) reference panel (938 unrelated individuals; downloaded from http://csg.sph.umich.edu/chaolong/LASER/)[15,16], using PLINK. We defined recent European ancestry as samples that fell into an ellipsoid spanning exclusively the European population within the HGDP panel. We restricted the analysis to individuals of recent European ancestry who passed quality control. Among these, 17,329 pairs of siblings comprising of 28,777 siblings, were inferred using KING[17], where an estimated kinship coefficient between 0.177 and 0.355, the proportion of the genomes that share two alleles IBD > 0.08, and the proportion of the genome that share zero alleles IBD > 0.04 corresponded to a full sibling pair.

HUNT descriptive data. There were 56,374 genotyped individuals in HUNT, including 11,448 families with at least two siblings comprising of 28,777 siblings (14,718 women) with complete data on genotype, height and education, diabetes and blood pressure (see supplementary figure 5 for a flow chart of participant inclusion and exclusion). On average the participants in the full unrelated sample were 48.7 (SD = 15.1) years old, had a BMI of 26.3 kg/m^2 (SD = 3.9), were 177.5 cm tall (SD = 6.6) and 164.3 cm tall (SD = 6.1) for men and women respectively, 2.5% of them had diabetes, and 42.5% had high blood pressure and had 12.3 years (SD = 2.3) of education. High blood pressure was defined as either currently taking anti-hypertensive medication or having systolic or diastolic blood pressure above 140 mmHg or 90 mmHg on average across up to three measurements in HUNT2.

Questionnaires, clinical measurements and hospitalizations. Participants attended a health survey which included comprehensive questionnaires, an interview and clinical examination. The participants' height and weight were measured with the participant wearing light clothes without shoes to the nearest centimetre and half kilogram, respectively. Education was defined using the question 'What is your highest level of education'. Participants answered one of five categories (1) primary school, (2) high school for 1 or 2 years, (3) complete high school, (4) college or university less than 4 years, and (5) college or university 4 years or more. Participants with university degrees were assigned to 16 years of education, those who completed high school were assigned 13 years, those who attended high school for 1 or 2 years were assigned 12 years, and those who only attended primary school were assigned 10 years. Diabetes was defined using responses to the question 'Have you had or do you have diabetes?', which has high validity[69]. High blood pressure was defined as those with systolic or diastolic blood pressure equal to or more than 140 or 90 mmHg, respectively, or reported use of antihypertensive medication.

Ethics. This study was approved by the Regional Committee for Medical and Health Research Ethics, Central Norway and all participants gave informed written consent (application numbers 2015/1209, 2015/2292 and 2017/2479).

The UK Biobank. The UK Biobank invited over 9 million people and sampled 503,317 participants from March 2006 to October 2010. The study sampled individuals from 21 study centres across Great Britain. A detailed description of the study can be read elsewhere[67,68,70]. The participants gave blood samples, from which DNA was extracted. Full details of the quality control process are available elsewhere[71]. Briefly, we excluded participants who had mismatched genetic and reported sex, or those with non-XX or XY chromosomes, extreme heterozygosity or missingness. We used variants in the Haplotype Reference Consortium (HRC) panel.

The UK Biobank descriptive data. There were up to 370,180 genotyped individuals in the UK Biobank, among whom were 16,847 families with at least two siblings, with 33,642 siblings (19,445 women) with complete data on genotype, height and education, diabetes and blood pressure. We restricted the analysis to siblings born in England to ensure that they experienced a similar school system. For a flow chart of participants inclusion and exclusion from the study see Supplementary Fig. 5. On average, the participants without siblings were 57.5 (SD = 7.4) years old, had a BMI of 27.4 kg m^{-2} (SD = 4.8), were 175.0 cm tall (SD = 6.8) and 162.8 cm tall (SD = 6.2) for men and women respectively, had 14.1 (SD = 2.3) years of education. 4.5% of them had diabetes and 54.0% had high blood pressure. High blood pressure was defined as either having a diagnosis of high blood

pressure or having systolic or diastolic blood pressure above 140 mmHg or 90 mmHg respectively on average across up to two clinic measurements.

Questionnaires, clinical measurements and hospitalizations. Weight (ID:21002) and standing height (ID:50) were measured using standardised instruments the baseline assessment centre visits. We defined education using the participants' response to the touch screen questionnaires about their educational qualifications (ID = 6138). We defined educational attainment using the participants' highest reported educational qualification at either measurement occasion. We assigned participants with university degrees to 17 years of education, those with professional qualifications such as teaching or nursing to 15 years, those with A-levels to 14 years, those with National Vocational Qualifications (NVQs), Higher National Diplomas (HNDs) to 13 years, General Certificate of Secondary Education (GCSEs), Certificate of Secondary Education (CSEs) or O-levels to 12 years, and those who reported no qualifications to 11 years, which was the legal minimum length of education for this cohort. Diabetes and high blood pressure were defined using responses to the self-reported touch screen questionnaire (ID = 6150 and ID = 2443). We used self-reported measures because measured blood pressure is affected by medication use. Missing values at the baseline visit were replaced by measures from subsequent visits if available.

Ethics. UK Biobank received ethical approval from the Research Ethics Committee (REC reference for UK Biobank is 11/NW/0382). This research was approved as part of application 8786.

23andMe replication. Individuals in the 23andMe replication dataset were customers of 23andMe, Inc., a personal genomics company. The 23andMe study protocols were approved by an external AAHRPP-accredited institutional review board and conducted in accordance with the Declaration of Helsinki principles.

There were 111,684 sibling pairs in the 23andMe dataset for a total of 223,368 genotyped individuals with complete data on genotype, height, BMI, education, diabetes, and blood pressure. For a flow chart of participants inclusion and exclusion from the study see Supplementary Fig. 6. Participants self-reported their mass (in kilograms) and height (in metres), from which BMI was calculated. To determine years of education, participants were asked 'What is the highest degree or level of school you have completed? If currently enrolled, please select the previous grade or highest degree received'. The following response options were then mapped onto the corresponding years of education: less than high school = 10 years, high school diploma or equivalency (GED) = 12 years, Associate's degree (for example, AA, AS) = 14 years, Vocational degree = 14 years, some college but no degree = 14 years, Bachelor's degree (for example, BA, BS) = 16 years, Master's degree (for example, MA, MS, MEng, MEd, MSW, MBA) = 19 years, Professional degree beyond a Bachelor's degree (for example, MD, DDS, DVM, LLB, JD) = 19 years, doctoral degree (for example, PhD, EdD) = 22 years. Research participants self-reported having ever been diagnosed or treated for both Type II diabetes or high blood pressure.

As previously described[73], DNA extraction and genotyping were performed on saliva samples by National Genetics Institute. Samples were genotyped on one of five Illumina-based genotyping platforms. Samples had minimum call rates of 98.5%. We phased participant data using either an internally developed tool, Finch (V1-V4 genotyping arrays) or Eagle2 (V5 genotyping array)[74]. We imputed phased research participant data using Minimac3 and a reference panel that combined both the May 2015 release of the 1000 Genomes Phase 3 haplotypes and the UK10K imputation reference panel ($n = 6285$). Throughout, we treated structural variants and small indels the same as SNPs. We used the same list of SNPs as for HUNT and UK Biobank, restricted to those that passed the 23andMe genotypic data QC.

We computed association test results using the sibling difference method assuming additive allelic effects logistic regression for case-control exposures, linear regression for quantitative exposures.

Genotype arrays. As previously described[73], samples were genotyped on one of five genotyping platforms. The v1 and v2 platforms were variants of the Illumina HumanHap550+ BeadChip, including about 25,000 custom SNPs selected by 23andMe, with a total of about 560,000 SNPs. The v3 platform was based on the Illumina OmniExpress+ BeadChip, with custom content to improve the overlap with our v2 array, with a total of about 950,000 SNPs. The v4 platform was a fully customized array, including a lower redundancy subset of v2 and v3 SNPs with additional coverage of lower-frequency coding variation, and about 570,000 SNPs. The v5 platform is an Illumina Infinium Global Screening Array (~640,000 SNPs) supplemented with ~50,000 SNPs of custom content.

The Finch phasing algorithm. As previously described[73], Finch implements the Beagle haplotype graph-based phasing algorithm, modified to separate the haplotype graph construction and phasing steps[75]. It extends the Beagle model to accommodate genotyping error and recombination, to handle cases where there are no consistent paths through the haplotype graph for the individual being phased. We constructed haplotype graphs for European and non-European samples on each 23andMe genotyping platform from a

representative sample of genotyped individuals, and then performed out-of-sample phasing of all genotyped individuals against the appropriate graph. For the X chromosome, we built separate haplotype graphs for the non-pseudo autosomal region and each pseudo autosomal region, and these regions were phased separately.

Imputation panel generation. As previously described[73], imputation panels created by combining multiple smaller panels have been shown to give better imputation performance than the individual constituent panels alone[76]. To that end, we combined the May 2015 release of the 1000 Genomes Phase 3 haplotypes with the UK10K imputation reference panel to create a single unified imputation reference panel[77,78]. To do this, multiallelic sites with N alternate alleles were split into N separate biallelic sites. We then removed any site whose minor allele appeared in only one sample. For each chromosome, we used Minimac3 to impute the reference panels against each other, reporting the best-guess genotype at each site[79]. This gave us calls for all samples over a single unified set of variants. We then joined these together to get, for each chromosome, a single VCF with phased calls at every site for 6285 samples.

Imputation. In preparation for imputation we split each chromosome of the reference panel into chunks of no more than 300,000 variants, with overlaps of 10,000 variants on each side. We used a single batch of 10,000 individuals to estimate Minimac3 imputation model parameters for each chunk. We imputed phased participant data against the chunked merged reference panel using Minimac3, treating males as homozygous pseudo-diploids for the non-pseudo autosomal region.

Selection of genotypes for instruments. For the analysis of HUNT and UK Biobank we selected 385 independent ($r^2 < 0.01$ within 10,000 kb) SNPs associated with height ($p < 5 \times 10^{-08}$) from Wood et al. and 79 associated with BMI in Locke et al.[65,72]. Neither HUNT nor UK Biobank were included in these studies. We clumped variants using the TwoSampleMR package[80]. We harmonized the alleles' effect sizes across the two samples and constructed weighted polygenic scores which were sums of the phenotype increasing alleles and weighted each variant by its effect on the phenotype in the published GWAS. For the analysis of the 23andMe data we used the subset of 347 and 64 SNPs available in the 23andMe data.

Empirical analyses. We compared seven empirical estimates of the effect of BMI on self-reported diabetes and high blood pressure and the effects of height and BMI on educational attainment. We used the familial fixed effects models (2) described above:

1. IPD ordinary least squares (OLS): The multivariable adjusted phenotypic association using ordinary least squares. Estimated using reg/plm commands.
2. IPD OLS family fixed effects (FE): The multivariable adjusted phenotypic association using ordinary least squares allowing for a family fixed effect across siblings. Estimated using xtreg/plm commands.
3. IPD MR-PRS unrelateds: A standard MR estimate of the effect of each exposure on the outcomes using the largest available sample of unrelated individuals. These models do not allow for any familial effects and are likely to suffer from bias. The estimate uses a polygenic score for each exposure and two stage least squares. Estimated separately in HUNT and UK Biobank using the ivreg and ivreg2 packages.
4. IPD MR-PRS siblings: As with 3. above, but restricted to siblings. Estimated using the ivreg2 package. We include this estimate to demonstrate the impact of including a family fixed effect on the estimates when holding the sample constant.
5. IPD MR-PRS siblings family fixed effects: An estimate using individual level data from the full sample of siblings allowing for family fixed effects. This allows for familial effects. Estimated separately in UK Biobank and HUNT using the xtivreg and plm packages[64]. This equivalent to family fixed effects models (vi and vii) described above and uses the same sample as 4. above.
6. 2SMR IVW siblings: An estimate using SNP summary data for Mendelian randomization including family fixed effects. The SNP-exposure and SNP-outcome associations were estimated on the same sample. The SNP level estimates of the effect of the exposure on the outcome were estimated separately in HUNT and UK Biobank and the overall Mendelian randomization (Wald) estimates are calculated for each SNP. For each SNP the Wald estimate is the ratio of the SNP-outcome and SNP-exposure association. We combine the estimate using random effects Inverse Variance Weighted (IVW) meta-analysis. This demonstrates how within family association estimates can be used within the two-sample MR framework.
7. 2SMR IVW siblings—split sample: As with 6. above, but the SNP-exposure and SNP-outcome associations were estimated in separate samples (i.e. split sample approach). The overall Wald estimates were combined via IVW meta-analysis as above. This ensures that there is no sample overlap between

the samples used to estimate the SNP-exposure and SNP-outcome associations. This eliminates the risk of weak instrument/sample overlap bias.

Covariates and standard errors. All analyses included age, sex, and the first 20 principal components of genetic variation. Cluster robust standard errors were used to allow for heteroskedasticity and allow for clustering and relatedness across siblings within families. Inclusion of the covariates age, sex, and principal components did not meaningfully affect the within family estimates, as they are independent of genotype conditional on sibling genotype. However, including these covariates may absorb some of the variation in the outcome and increase the precision of our estimates.

Sensitivity analyses. Finally, we tested for difference ($p_{\text{diff unrelated}}$) between the Mendelian randomization estimates using the unrelated individuals and the summary data within-family estimates using the split sample approach (i.e. as in 6. above)[81]. We investigated whether our results could be explained by pleiotropy using the weighted median, weighted modal and MR-Egger estimators and the SNP-phenotype associations allowing for a family fixed effect[8,10,82]. We used a split sample approach in which the SNP-exposure and SNP-outcome associations were estimated in separate samples. We estimated the percentage change in the SNP-phenotype coefficients with and without allowing for a family fixed effect.

Shrinkage. We investigated shrinkage of the total to within family SNP-phenotype associations using seemingly unrelated regression. We estimated the shrinkage for each of the 455 SNPs included in the analysis and the five phenotypes (education, BMI, height, diabetes and high blood pressure). We then meta-analysed estimates for each SNP and phenotype across the two studies. Finally, independently for each phenotype we meta-analysed across all 455 SNPs used in the analysis to give an average shrinkage for all SNPs.

Data availability

Data from the HUNT study was accessed under ethics approvals 2015/1209 REK midt 2015/2292 REK midt, and 2017/2479 REK sør-øst, and project number 2019/2181. Data from the UK Biobank was accessed as part of application 8786. The empirical datasets used with the HUNT study and UK Biobank will be archived with the studies and will be made available to individuals who obtain the necessary permissions from the studies' data access committees. Data from 23andMe was processed by 23andMe, and the individual level data cannot be made publicly available, however, 23andMe do provide access to summary data via a system of managed access. If you would like to apply for access, please see the following website for more details https://research.23andme.com/dataset-access/.

Code availability

The code used to clean and analyse the data and the SNP level summary statistics are available here: https://github.com/nmdavies/within_family_mr. Code for the simulations are available here: https://github.com/mrcieu/mrtwin_power.

References

1. Davey Smith, G. & Ebrahim, S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *Int. J. Epidemiol.* **32**, 1–22 (2003).
2. Davey Smith, G. & Hemani, G. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Hum. Mol. Genet.* **23**, R89–98 (2014).
3. Davies, N. M., Holmes, M. V. & Davey Smith, G. Reading Mendelian randomisation studies: a guide, glossary, and checklist for clinicians. *BMJ.* https://doi.org/10.1136/bmj.k601 (2018).
4. Pingault, J.-B. et al. Using genetic data to strengthen causal inference in observational research. *Nat. Rev. Genet.* **19**, 566–580 (2018).
5. Katan, M. Apoupoprotein E isoforms, serum cholesterol, and cancer. *Lancet* **327**, 507–508 (1986).
6. Hartwig, F. P., Davies, N. M. & Davey Smith, G. Bias in Mendelian randomization due to assortative mating. *Genet. Epidemiol.* **42**, 608–620 (2018).
7. Windmeijer, F., Farbmacher, H., Davies, N. & Davey Smith, G. On the use of the lasso for instrumental variables estimation with some invalid instruments. *J. Am. Stat. Assoc.* https://doi.org/10.1080/01621459.2018.1498346 (2018).
8. Hartwig, F. P., Davey Smith, G. & Bowden, J. Robust inference in summary data Mendelian randomization via the zero modal pleiotropy assumption. *Int. J. Epidemiol.* **46**, 1985–1998 (2017).
9. Bowden, J., Davey Smith, G. & Burgess, S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int. J. Epidemiol.* **44**, 512–525 (2015).
10. Bowden, J., Davey Smith, G., Haycock, P. C. & Burgess, S. Consistent estimation in Mendelian randomization with some invalid instruments using a weighted median estimator. *Genet. Epidemiol.* **40**, 304–314 (2016).
11. Burgess, S. & Thompson, S. G. Multivariable Mendelian randomization: the use of pleiotropic genetic variants to estimate causal effects. *Am. J. Epidemiol.* **181**, 251–260 (2015).
12. Mendel, G. Experiments in plant hybridization. Verhandlungen des naturforschenden Vereines Brünn. www.mendelweb.org/Mendel.html (1865).
13. Haworth, S. et al. Apparent latent structure within the UK Biobank sample has implications for epidemiological analysis. *Nat. Commun.* **10**, 333 (2019).
14. Kong, A. et al. The nature of nurture: effects of parental genotypes. *Science* **359**, 424–428 (2018).
15. Robinson, M. R. et al. Genetic evidence of assortative mating in humans. *Nat. Hum. Behav.* **1**, 0016 (2017).
16. Beauchamp, J. P., Cesarini, D., Johannesson, M., Lindqvist, E. & Apicella, C. On the sources of the height–intelligence correlation: New insights from a bivariate ACE model with assortative mating. *Behav. Genet.* **41**, 242–252 (2011).
17. Fisher, R. The correlation between relatives on the supposition of Mendelian inheritance. *Trans. R. Soc. Edinb.* **52**, 399–433 (1918).
18. Howe, L. J. et al. Genetic evidence for assortative mating on alcohol consumption in the UK Biobank. *Nat. Commun.* **10**, 5039 (2019).
19. Nordsletten, A. E. et al. Patterns of nonrandom mating within and across 11 major psychiatric disorders. *JAMA Psychiatry* **73**, 354 (2016).
20. Silventoinen, K., Kaprio, J., Lahelma, E., Viken, R. J. & Rose, R. J. Assortative mating by body height and BMI: finnish twins and their spouses. *Am. J. Hum. Biol.* **15**, 620–627 (2003).
21. Mathieson, I. & McVean, G. Differential confounding of rare and common variants in spatially structured populations. *Nat. Genet.* **44**, 243–246 (2012).
22. Berg, J. J. et al. Reduced signal for polygenic adaptation of height in UK Biobank. *eLife* **8**, e39725 (2019).
23. Fulker, D. W., Cherny, S. S., Sham, P. C. & Hewitt, J. K. Combined linkage and association sib-pair analysis for quantitative traits. *Am. J. Hum. Genet.* **64**, 259–267 (1999).
24. Spielman, R. S. & Ewens, W. J. A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test. *Am. J. Hum. Genet.* **62**, 450–458 (1998).
25. Warrington, N. M., Freathy, R. M., Neale, M. C. & Evans, D. M. Using structural equation modelling to jointly estimate maternal and fetal effects on birthweight in the UK Biobank. *Int. J. Epidemiol.* https://doi.org/10.1093/ije/dyy015 (2018).
26. Abecasis, G. R., Cardon, L. R. & Cookson, W. O. C. A general test of association for quantitative traits in nuclear families. *Am. J. Hum. Genet.* **66**, 279–292 (2000).
27. Weiner, D. J. et al. Polygenic transmission disequilibrium confirms that common and rare variation act additively to create risk for autism spectrum disorders. *Nat. Genet.* **49**, 978–985 (2017).
28. Spielman, R. S. & Ewens, W. J. The TDT and other family-based tests for linkage disequilibrium and association. *Am. J. Hum. Genet.* **59**, 983–989 (1996).
29. 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.* **50**, 1112–1121 (2018).
30. Okbay, A. et al. Genome-wide association study identifies 74 loci associated with educational attainment. *Nature* **533**, 539–542 (2016).
31. Hemani, G. et al. Inference of the genetic architecture underlying BMI and height with the use of 20,240 sibling pairs. *Am. J. Hum. Genet.* **93**, 865–875 (2013).
32. Minică, C. C., Dolan, C. V., Boomsma, D. I., de Geus, E. & Neale, M. C. Extending causality tests with genetic instruments: an integration of Mendelian randomization with the classical twin design. *Behav. Genet.* **48**, 337–349 (2018).
33. Heath, A. C. et al. Testing hypotheses about direction of causation using cross-sectional family data. *Behav. Genet* **23**, 29–50 (1993).
34. Kang, H., Zhang, A., Cai, T. T. & Small, D. S. Instrumental variables estimation with some invalid instruments and its application to Mendelian randomization. *J. Am. Stat. Assoc.* **111**, 132–144 (2016).
35. Lawson, D. J. et al. Is population structure in the genetic biobank era irrelevant, a challenge, or an opportunity? *Hum. Genet.* https://doi.org/10.1007/s00439-019-02014-8 (2019).
36. Angrist, J. D. & Krueger, A. B. Split-sample instrumental variables estimates of the return to schooling. *J. Bus. Econ. Stat.* **13**, 225–235 (1995).

37. Pierce, B. L. & Burgess, S. Efficient design for mendelian randomization studies: subsample and 2-sample instrumental variable estimators. *Am. J. Epidemiol.* **178**, 1177–1184 (2013).

38. Neale, M. C. et al. Distinguishing population stratification from genuine allelic effects with Mx: association of ADH2 with alcohol consumption. *Behav. Genet.* **29**, 233–243 (1999).

39. Neale, B. M., Ferreira, M., Medland, S. E. & Posthuma, D. *Statistical genetics: gene mapping through linkage and association*. (Taylor & Francis Group, 2008).

40. Ott, J., Kamatani, Y. & Lathrop, M. Family-based designs for genome-wide association studies. *Nat. Rev. Genet.* **12**, 465–474 (2011).

41. Holmes, M. V. et al. Causal effects of body mass index on cardiometabolic traits and events: a mendelian randomization analysis. *Am. J. Hum. Genet.* **94**, 198–208 (2014).

42. Dolan, C. V., de Kort, J. M., van Beijsterveldt, T. C. E. M., Bartels, M. & Boomsma, D. I. GE covariance through phenotype to environment transmission: an assessment in longitudinal twin data and application to childhood anxiety. *Behav Genet* **44**, 240–253 (2014).

43. Moscati, A., Verhulst, B., McKee, K., Silberg, J. & Eaves, L. Cross-lagged analysis of interplay between differential traits in sibling pairs: validation and application to parenting behavior and ADHD symptomatology. *Behav. Genet.* **48**, 22–33 (2018).

44. Baud, A. et al. Genetic variation in the social environment contributes to health and disease. *PLoS Genet.* **13**, e1006498 (2017).

45. Magnusson, P. K. E., Rasmussen, F. & Gyllensten, U. B. Height at age 18 years is a strong predictor of attained education later in life: cohort study of over 950,000 Swedish men. *Int. J. Epidemiol.* **35**, 658–663 (2006).

46. Case, A. & Paxson, C. Stature and status: Height, ability, and labor market outcomes. *J. Polit. Econ.* **116**, 499–532 (2008).

47. Lundborg, P., Nystedt, P. & Rooth, D.-O. Body size, skills, and income: evidence from 150,000 teenage siblings. *Demography* **51**, 1573–1596 (2014).

48. Silventoinen, K., Kaprio, J. & Lahelma, E. Genetic and environmental contributions to the association between body height and educational attainment: a study of adult Finnish twins. *Behav. Genet.* **30**, 477–485 (2000).

49. Silventoinen, K., Krueger, R. F., Bouchard, T. J., Kaprio, J. & McGue, M. Heritability of body height and educational attainment in an international context: Comparison of adult twins in Minnesota and Finland. *Am. J. Hum. Biol.* **16**, 544–555 (2004).

50. Tyrrell, J. et al. Height, body mass index, and socioeconomic status: Mendelian randomisation study in UK Biobank. *BMJ* i582 (2016), https://doi.org/10.1136/bmj.i582.

51. Böckerman, P. et al. The effect of weight on labor market outcomes: An application of genetic instrumental variables. *Health Econ.* **28**, 65–77 (2019).

52. Davey Smith, G. & Davies, N. M. Can genetic evidence help us understand why height and weight relate to social position? *BMJ* https://doi.org/10.1136/bmj.i1224. (2016).

53. DiPrete, T. A., Burik, C. A. P. & Koellinger, P. D. Genetic instrumental variable regression: Explaining socioeconomic and health outcomes in nonexperimental data. *Proc. Natl Acad. Sci. USA* **115**, E4970–E4979 (2018).

54. Davies, N. M. et al. Within family Mendelian randomization studies. *Hum. Mol. Genet.* **28**, R170–R179 (2019).

55. Hwang, L.-D., Davies, N. M., Warrington, N. M. & Evans, D. M. Integrating family-based and mendelian randomization designs. *Cold Spring Harb. Perspect. Med.* https://doi.org/10.1101/cshperspect.a039503 (2020).

56. Carey, G. Sibling imitation and contrast effects. *Behav. Genet.* **16**, 319–341 (1986).

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

58. Saudino, K. J., Wertz, A. E., Gagne, J. R. & Chawla, S. Night and day: are siblings as different in temperament as parents say they are? *J. Pers. Soc. Psychol.* **87**, 698–706 (2004).

59. Simonoff, E. et al. Genetic influences on childhood hyperactivity: contrast effects imply parental rating bias, not sibling interaction. *Psychol. Med.* **28**, 825–837 (1998).

60. Moen, G.-H., Hemani, G., Warrington, N. M. & Evans, D. M. Calculating Power to detect maternal and offspring genetic effects in genetic association studies. *Behav. Genet.* https://doi.org/10.1007/s10519-018-9944-9. (2019).

61. Evans, D. M., Moen, G.-H., Hwang, L.-D., Lawlor, D. A. & Warrington, N. M. Elucidating the role of maternal environmental exposures on offspring health and disease using two-sample Mendelian randomization. *Int. J. Epidemiol.* https://doi.org/10.1093/ije/dyz019. (2019).

62. Hartwig, F. P. & Davies, N. M. Why internal weights should be avoided (not only) in MR-Egger regression. *Int. J. Epidemiol.* https://doi.org/10.1093/ije/dyw240. (2016).

63. Bowden, J. et al. Improving the accuracy of two-sample summary-data Mendelian randomization: moving beyond the NOME assumption. *Int. J. Epidemiol.* **48**, 728–742 (2019).

64. Mark E. Schaffer. XTIVREG2: Stata module to perform extended IV/2SLS, GMM and AC/HAC, LIML and k-class regression for panel data models. (2005).

65. Locke, A. E. et al. Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518**, 197–206 (2015).

66. Krokstad, S. et al. Cohort Profile: The HUNT Study, Norway. *Int. J. Epidemiol.* **42**, 968–977 (2013).

67. Bycroft, C. et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).

68. Fry, A. et al. Comparison of sociodemographic and health-related characteristics of UK Biobank participants with those of the general population. *Am. J. Epidemiol.* **186**, 1026–1034 (2017).

69. Midthjell, K., Holmen, J., Bjørndal, A. & Lund-Larsen, G. Is questionnaire information valid in the study of a chronic disease such as diabetes? The Nord-Trøndelag diabetes study. *J. Epidemiol. Community Health* **46**, 537–542 (1992).

70. Sudlow, C. et al. UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* **12**, e1001779 (2015).

71. Mitchell, R., Hemani, G., Dudding, T. & Paternoster, L. UK Biobank Genetic Data: MRC-IEU Quality Control, Version 1. (2017) https://doi.org/10.5523/bris.3074krb6t2frj29yh2b03x3wxj.

72. Wood, A. R. et al. Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat. Genet.* **46**, 1173–1186 (2014).

73. Heilbron, K. et al. Unhealthy Behaviours and Parkinsons Disease: A Mendelian Randomisation Study. https://doi.org/10.1101/2020.03.25.20039230 http://medrxiv.org/lookup/doi/10.1101/2020.03.25.20039230 (2020).

74. Loh, P.-R. et al. Reference-based phasing using the Haplotype Reference Consortium panel. *Nat. Genet.* **48**, 1443–1448 (2016).

75. Browning, S. R. & Browning, B. L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.* **81**, 1084–1097 (2007).

76. UK10K Consortium. et al. Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. *Nat. Commun.* **6**, 8111 (2015).

77. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

78. The UK10K Consortium. The UK10K project identifies rare variants in health and disease. *Nature* **526**, 82–90 (2015).

79. Das, S. et al. Next-generation genotype imputation service and methods. *Nat. Genet.* **48**, 1284–1287 (2016).

80. Hemani, G. et al. The MR-Base platform supports systematic causal inference across the human phenome. *eLife.* **7**, e34408 (2018).

81. Altman, D. G. & Bland, J. M. Interaction revisited: the difference between two estimates. *BMJ* **326**, 219 (2003).

82. Bowden, J. et al. Assessing the suitability of summary data for two-sample Mendelian randomization analyses using MR-Egger regression: the role of the I2 statistic. Int. J. Epidemiol. 45, 1961–1974 (2016). *Int. J. Epidemiol.* **45**, 1961–1974 (2016).

Acknowledgements

Author contributions

G.D.S., N.M.D., K.H. and B.M.B. obtained funding for this study. B.M.B., G.H. and N.M.D. designed, analysed and cleaned the data, interpreted results, wrote and revised the paper. G.H. ran the simulation study. E.S. provided the proof comparing the difference and fixed effects estimators, interpreted results and revised the paper. K.H. ran the replication in 23andMe data, drafted the sections relating to the 23andMe data and revised the paper. G.D.S. conceived of the study, interpreted the results, wrote and revised the paper. F.P.H., S.H., G.A.V., Y.C., L.D.H., A.H., D.I.B., A.H., J.H., M.N., M.G.N., N.L.P., C.A.R., E.M.T.-D., A.G., L.H., T.M., S.L., A.A., F.W., W-M.C., J.H.B., K.H., C.W., D.M.E., J.K. and B.O.A. interpreted the results and revised the paper.

Competing interests

Additional information

Ben Brumpton [1,2,3✉], Eleanor Sanderson [2,4], Karl Heilbron[5], Fernando Pires Hartwig [2,6], Sean Harrison [2,4], Gunnhild Åberge Vie [1], Yoonsu Cho [2,4], Laura D. Howe[2,4], Amanda Hughes[2,4], Dorret I. Boomsma[7], Alexandra Havdahl [2,8,9], John Hopper[10], Michael Neale [11], Michel G. Nivard [7], Nancy L. Pedersen[12], Chandra A. Reynolds[13], Elliot M. Tucker-Drob[14], Andrew Grotzinger [14], Laurence Howe[2,4], Tim Morris [2,4], Shuai Li[10,15], The Within-family Consortium*, The 23andMe Research Team*, Adam Auton[5], Frank Windmeijer[2,16], Wei-Min Chen[17], Johan Håkon Bjørngaard[1,18], Kristian Hveem[1], Cristen Willer [19,20,21], David M. Evans [2,22], Jaakko Kaprio [23,24], George Davey Smith [2,4,26], Bjørn Olav Åsvold[1,25,26], Gibran Hemani [2,4,26] & Neil M. Davies [1,2,4,26✉]

[1]K.G. Jebsen Center for Genetic Epidemiology, Department of Public Health and Nursing, NTNU, Norwegian University of Science and Technology, Trondheim, Norway. [2]Medical Research Council Integrative Epidemiology Unit, University of Bristol, BS8 2BN Bristol, UK. [3]Clinic of Thoracic and Occupational Medicine, St. Olavs Hospital, Trondheim University Hospital, Trondheim, Norway. [4]Population Health Sciences, Bristol Medical School, University of Bristol, Barley House, Oakfield Grove, Bristol BS8 2BN, UK. [5]23andMe, Inc., 223 N Mathilda Avenue, Sunnyvale, CA 94086, USA. [6]Postgraduate Program in Epidemiology, Federal University of Pelotas, Pelotas, Brazil. [7]Netherlands Twin Register, Department of Biological Psychology, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands. [8]Nic Waals Institute, Lovisenberg Diaconal Hospital, Spångbergveien 25, 0853 Oslo, Norway. [9]Department of Mental Disorders, Norwegian Institute of Public Health, Sandakerveien 24 C, 0473 Oslo, Norway. [10]Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, The University of Melbourne, 207 Bouverie Street, Carlton, VIC 3053, Australia. [11]Virginia Institute for Psychiatric and Behavior Genetics, Virginia Commonwealth University, Richmond, VA, USA. [12]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden. [13]Department of Psychology, University of California Riverside, Riverside, CA, USA. [14]Department of Psychology and Population Research Center, University of Texas at Austin, 108 E. Dean Keeton Stop A8000,, Austin, TX 78712, USA. [15]Centre for Cancer Genetic Epidemiology, Department of Public Health and Primary Care, University of Cambridge, Strangeways Research Laboratory, Worts Causeway, Cambridge CB1 8RN, UK. [16]Department of Economics, University of Bristol, 2 Priory Road, Bristol BS8 1TU, UK. [17]Center for public health genomics, Department of public health sciences, University of Virginia, Charlottesville, VA, USA. [18]Faculty of Nursing and Health Sciences, Nord University, Levanger, Norway. [19]Department of Biostatistics and Center for Statistical Genetics, University of Michigan, Ann Arbor, MI, USA. [20]Department of Internal Medicine, University of Michigan, Ann Arbor, MI, USA. [21]Department of Human Genetics, University of Michigan, Ann Arbor, MI, USA. [22]University of Queensland Diamantina Institute, University of Queensland,

Brisbane, QLD, Australia. [23]Department of Public Health, University of Helsinki, Helsinki, Finland. [24]Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland. [25]Department of Endocrinology, St Olavs Hospital, Trondheim University Hospital, Trondheim, Norway. [26]These authors contributed equally: George Davey Smith, Bjørn Olav Åsvold, Gibran Hemani, Neil M. Davies. *Lists of authors and their affiliations appear at the end of the paper. ✉email: ben.brumpton@ntnu.no; neil.davies@bristol.ac.uk

The Within-family Consortium

Ben Brumpton[1,2,3], Eleanor Sanderson[2,4], Karl Heilbron[5], Fernando Pires Hartwig[2,6], Sean Harrison[2,4], Gunnhild Åberge Vie[1], Yoonsu Cho[2,4], Laura D. Howe[2,4], Amanda Hughes[2,4], Dorret I. Boomsma[7], Alexandra Havdahl[2,8,9], John Hopper[10], Michael Neale[11], Michel G. Nivard[7], Nancy L. Pedersen[12], Chandra A. Reynolds[13], Elliot M. Tucker-Drob[14], Andrew Grotzinger[14], Laurence Howe[2,4], Tim Morris[2,4], Shuai Li[10,15], Adam Auton[5], Frank Windmeijer[2,16], Wei-Min Chen[17], Johan Håkon Bjørngaard[1,18], Kristian Hveem[1], Cristen Willer[19,20,21], David M. Evans[2,22], Jaakko Kaprio[23,24], George Davey Smith[2,4], Bjørn Olav Åsvold[1,25], Gibran Hemani[2,4] & Neil M. Davies[1,2,4]

The 23andMe Research Team

Karl Heilbron[5] & Adam Auton[5]

A full list of consortium members is included in the Supplementary materials.