

# Assessing causal links between age at menarche and adolescent mental health: a Mendelian randomisation study

Adrian Dahl Askelund<sup>1,2\*</sup> ([jaaskelu@uio.no](mailto:jaaskelu@uio.no)), Robyn E Wootton<sup>2,3</sup> ([robyn.wootton@bristol.ac.uk](mailto:robyn.wootton@bristol.ac.uk)),  
Fartein A Torvik<sup>1,4</sup> ([fartein.ask.torvik@fhi.no](mailto:fartein.ask.torvik@fhi.no)), Rebecca B Lawn<sup>5</sup> ([rlawn@hsph.harvard.edu](mailto:rlawn@hsph.harvard.edu)),  
Helga Ask<sup>6</sup> ([helga.ask@fhi.no](mailto:helga.ask@fhi.no)), Elizabeth Corfield<sup>6</sup> ([elizabeth.corfield@fhi.no](mailto:elizabeth.corfield@fhi.no)), Maria C Magnus<sup>3,4</sup>  
([mariachristine.magnus@fhi.no](mailto:mariachristine.magnus@fhi.no)), Ted Reichborn-Kjennerud<sup>6,7</sup> ([ted.reichborn-kjennerud@fhi.no](mailto:ted.reichborn-kjennerud@fhi.no)),  
Per M Magnus<sup>4</sup> ([perminor.magnus@fhi.no](mailto:perminor.magnus@fhi.no)), Ole A Andreassen<sup>8,9</sup> ([ole.andreassen@medisin.uio.no](mailto:ole.andreassen@medisin.uio.no)),  
Camilla Stoltenberg<sup>13,14</sup> ([camilla.stoltenberg@fhi.no](mailto:camilla.stoltenberg@fhi.no)), George Davey Smith<sup>3</sup>  
([kz.davey-smith@bristol.ac.uk](mailto:kz.davey-smith@bristol.ac.uk)), Neil M Davies<sup>3,10,11</sup> ([neil.davies@bristol.ac.uk](mailto:neil.davies@bristol.ac.uk)),  
Alexandra Havdahl<sup>2,3,6,12</sup> ([alexandra.havdahl@psykologi.uio.no](mailto:alexandra.havdahl@psykologi.uio.no)),  
& Laurie J Hannigan<sup>2,3,6\*</sup> ([laurie.hannigan@bristol.ac.uk](mailto:laurie.hannigan@bristol.ac.uk))

<sup>1</sup>Department of Psychology, University of Oslo, Oslo, Norway

<sup>2</sup>Nic Waals Institute, Lovisenberg Diaconal Hospital, Oslo, Norway

<sup>3</sup>MRC Integrative Epidemiology Unit, Bristol Medical School, University of Bristol, Bristol, UK

<sup>4</sup>Centre for Fertility and Health, Norwegian Institute of Public Health, Oslo, Norway

<sup>5</sup>Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, US

<sup>6</sup>Department of Mental Disorders, Norwegian Institute of Public Health, Oslo, Norway

<sup>7</sup>Institute of Clinical Medicine, University of Oslo, Oslo, Norway

<sup>8</sup>NORMENT Centre, Institute of Clinical Medicine, University of Oslo, Oslo, Norway

<sup>9</sup>Division of Mental Health and Addiction, Oslo University Hospital, Oslo, Norway

<sup>10</sup>Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, UK

<sup>11</sup>KG Jebsen Center for Genetic Epidemiology, Department of Public Health and Nursing, Norwegian University of Science and Technology, Trondheim, Norway

<sup>12</sup>Promenta Research Center, Department of Psychology, University of Oslo, Oslo, Norway

<sup>13</sup>Norwegian Institute of Public Health, Oslo, Norway

<sup>14</sup>Department of Global Public Health and Primary Care, University of Bergen, Bergen, Norway

\*Corresponding authors: Adrian Dahl Askelund ([jaaskelu@uio.no](mailto:jaaskelu@uio.no)) / Laurie J Hannigan ([laurie.hannigan@bristol.ac.uk](mailto:laurie.hannigan@bristol.ac.uk))

<b>Abstract</b>	3
<b>Background</b>	4
<b>Methods</b>	8
Design	8
Sample	8
Measures	9
Statistical analysis	13
Observational analyses	13
Mendelian randomisation analyses	14
Equivalence testing	18
Negative control analyses	19
Sampling plan	19
Inclusion criteria and sample size	19
Power calculations	21
Analysis plan	24
Question 1: To what extent is age at menarche associated with adolescent depression?	25
Question 2: Does age at menarche associate with symptoms or diagnoses in other domains of mental health, independent of depression?	27
Question 3: What is the evidence for a causal link between age at menarche and depression?	28
Question 4: Is there evidence of causal links between age at menarche and other domains of mental health?	32
<b>Abbreviations</b>	34
<b>Declarations</b>	35
Ethics approval and consent to participate	35
Consent for publication	36
Availability of data and materials	36
Competing interests	36
Funding	36
Author's contributions	37
Acknowledgements	37

## Abstract

The timing of puberty may have a substantial impact on adolescent mental health. In particular, earlier age at menarche has been associated with elevated rates of depression in adolescents. Some preliminary evidence suggests that this relationship is causal, but replication and an investigation of the broader generalisability of this finding is warranted. In this Registered Report, we will triangulate different causal inference methods using a new wave of data from the Norwegian Mother, Father and Child Cohort Study (MoBa). We will investigate: (1) to what extent age at menarche is associated with adolescent depression; (2) if associations extend to other domains of mental health; and (3) whether links between age at menarche and mental health are likely to be causal. The findings will elucidate the impact of pubertal timing on mental health and may help inform intervention development.

*Keywords: puberty, age at menarche, adolescent mental health, depression, causal inference, mendelian randomisation, registered report*

## Background

Early pubertal timing has been associated with problems in a wide range of adolescent mental health domains (e.g., depression [1–13], anxiety [7, 14, 15], conduct disorders [3, 7, 16–19], and attention-deficit hyperactivity disorder; ADHD [6]) across different indicators of pubertal development and across sexes [20]. The consistency of associations between early timing and adolescent mental health has led to the hypothesis that early pubertal timing is a transdiagnostic risk factor for psychopathology in adolescents [21].

Despite the apparent generality of associations between early pubertal timing and adolescent mental health, the prominent rise in rates of female depression beginning during puberty [22] has led to this outcome receiving particular empirical focus [23]. Timing of puberty in females is commonly indexed using onset of menses (menarche). Earlier age at menarche has been associated with elevated depressive symptoms in adolescents in several observational studies [4, 24–32], but not in all [33–37], and also higher rates of clinical depression during adolescence [31, 37]. However, although early pubertal maturation in females has been associated with a wide range of problems in adolescence, these associations may dissipate by adulthood [3, 38]. A notable exception in a large prospective study was that heightened risk of depression persisted into young adulthood for early maturers, in particular for those with a history of conduct disorder [3].

The association between pubertal timing and depression in adolescent females may be due to the biological underpinnings of reproductive maturation. The female sex hormone estradiol increases with puberty and is associated with depression [39, 40], and hormonal contraceptive use has been associated with higher levels of depressive symptoms, especially in adolescence [41]. In fact, it has been found that stage of breast development (governed primarily by estradiol) was associated with depression independently of timing of menarche in adolescent females from the Avon Longitudinal Study of Parents and Children

(ALSPAC) [42]. Interestingly, a recent study in the same sample found that a polygenic score for age at menarche showed a potential indirect association with adolescent depressive symptoms through stage of breast development [43]. Alongside psychosocial pathways (i.e., visible breast development leading to unwanted sexual attention at a younger age), increases in estradiol represents a plausible biological mechanism for the link between early pubertal timing and depression in females.

Despite the series of observational studies, it is unknown whether the link between age at menarche and depression represents a truly causal relationship. This is important because several robust observational associations in epidemiology have turned out not to be causal and may instead have been the result of confounding (i.e., vitamin E supplement use and cardiovascular disease [44]). In the case of associations between age at menarche and depression, body mass index (BMI) is a particularly likely candidate for confounding given the robust (and plausibly causal) links between BMI and age at menarche [45] and between BMI and depression [46–48]. Failure to appropriately account for potential confounding, especially by BMI [4, 25, 27–30, 32–37], has been a relatively common shortcoming of the literature on this topic to date. In previous studies that explicitly controlled for BMI, the relationship was somewhat attenuated [26, 49]. However, another study also found BMI to be a partial mediator of the relationship between earlier menarche and depression [31].

Mendelian randomisation (MR) is a causal inference method that can be implemented in instrumental variable analyses [50] which is particularly useful when experimental manipulation of the variable of interest is not ethical or feasible. Since hundreds of genetic variants are strongly linked with age at menarche [51], single nucleotide polymorphisms (SNPs) that are independently associated with this phenotype can be used as genetic instruments in MR analyses. The logic of MR is analogous to that of a randomised controlled trial (RCT). Unlike in an RCT design where individuals are randomly assigned to experimental groups, in MR we use random “assignment” to genotype (ensured by the

random transmission of one of two possible alleles at each genetic locus from each parent to their child at conception [52]). Specifically, these genetic variants are used as instrumental variables, serving as a genetic proxy for age at menarche.

Whereas self-reported age at menarche may be associated with several different confounders (even if precisely measured), the genetic instrument is assumed to be independent of such confounding. Both the widespread genetic influence on age at menarche [51] and the high accuracy and reliability of self-reported age at menarche [53] jointly increase the strength of the genetic instrument employed here, which serves to improve study power and minimise weak instrument bias [54]. The strength of the genetic instrument makes MR especially valuable for advancing menarche research. Provided that some important assumptions of MR hold true, we can estimate the causal effects of age at menarche on adolescent mental health.

A previous study found preliminary evidence that the relationship between age at menarche and depression in early adolescence may be causal, using MR in ALSPAC ( $N = 2404$ ) [55]. Specifically, they found that early age at menarche resulted in more depressive symptoms at age 14 (independent of BMI), but not later in adolescence. However, this study had low power due to a modest sample size for MR. Here, we aim to replicate the 14-year analyses in adolescents from a larger birth cohort, the Norwegian Mother, Father and Child Cohort Study (MoBa) [56]. This replication will allow for a confirmatory and higher-powered test of the hypothesis that earlier age at menarche is causally related to adolescent depression.

Beyond replicating its key finding, we will also extend the previous approach [55] in several key ways. First, we will test whether effects of earlier age at menarche extend to other domains of mental health (anxiety disorders, conduct disorder (CD), oppositional defiant disorder (ODD), and ADHD), independent of associations with depression. Second, we will use multivariable methods to examine different confounders or mechanisms, by

simultaneously including genetic instruments for childhood body size, adult BMI or estradiol in the MR model together with age at menarche. Third, in line with recommendations to triangulate evidence across approaches for robust causal inference [57], we will combine MR with negative control analyses using symptoms prior to puberty as a negative control outcome. This triangulation is particularly important in the context of replication studies, given that the same sources of bias could lead to results being replicated in another study using the same methodology [58, 59].

A previous hypothesis-free MR phenome-wide association study identified potential causal effects of age at menarche on adult mental health [60], but these were not followed up with replication in any independent cohorts. Here we take a confirmatory approach, testing causal hypotheses about the role of age at menarche in the aetiology of developing mental health disorders. This is important in part because a causal effect of age at menarche may help explain the sharp rise in depression rates among females from early adolescence [22]. This research might further help with identifying female adolescents at increased risk, facilitating early identification and prevention of mental health problems in adolescence and beyond.

To test our hypotheses we make use of the Registered Report format, demonstrating its applicability to epidemiological analyses of cohort data when a new wave of data collection ensures that the exposure and outcome data has not been observed prior to the analytic choices being made. This format, combined with several sensitivity tests, will strengthen our statistical inferences by preserving false positive rates at the specified level [61] and ultimately increase confidence in causal conclusions that are drawn.

In summary, we will address the following questions and hypotheses (detailed overview in Additional file 1):

1. To what extent is age at menarche associated with adolescent depression?

- a. We hypothesise that earlier age at menarche will be associated with elevated depressive symptoms at age 14 (H1a)
  - b. We hypothesise that earlier age at menarche will be associated with higher rates of depression diagnoses during adolescence (H1b)
2. Does age at menarche associate with symptoms or diagnoses in other domains of mental health (anxiety, CD, ODD, or ADHD), independent of depression?
  - a. We hypothesise that age at menarche will be associated with symptoms in other domains at age 14, independent of depressive symptoms (H2.1-4a)
  - b. We hypothesise that age at menarche will be associated with diagnoses in other domains in adolescence, independent of depressive disorders (H2.1-3b)
3. What is the evidence for a causal link between age at menarche and depression?
  - a. We hypothesise that earlier age at menarche will show a causal relationship, resulting in elevated depressive symptoms at age 14 (H3a)
  - b. We hypothesise that earlier age at menarche will show a causal relationship, resulting in higher rates of depression diagnoses during adolescence (H3b)
4. Is there evidence of causal links between age at menarche and other domains of mental health?
  - a. We hypothesise that age at menarche will show a causal relationship with symptoms in other domains (H4.1-4a)
  - b. We hypothesise that age at menarche will show a causal relationship with rates of diagnoses in other domains (H4.1-3b)

## Methods

### Design

#### *Sample*

The Norwegian Mother, Father and Child Cohort Study is a population-based pregnancy cohort study conducted by the Norwegian Institute of Public Health [56]. Pregnant women and their partners were recruited at approximately pregnancy week 17 between 1999 and 2008. The women consented to participation in 41% of the pregnancies. The cohort now includes 114,500 children, 95,200 mothers and 75,200 fathers.



In MoBa, phenotype data have been collected by questionnaires from early pregnancy to middle childhood, provided primarily by mothers (around week 17, 22 and 30 of pregnancy, when the child was 6 and 18 months, and at 3, 5, and 8 years). This project will also make use of an ongoing wave of data collection in adolescence (questionnaires returned at ~ 14.5; hereafter age 14). The 14-year data were not available to us during the preparation of the stage 1 element of the Registered Report.

## **Measures**

*Exposures.* Self-reported age at menarche (in years) from the 14-year questionnaire will be included as the main exposure. We will run the observational analyses with both a continuous and a categorical (early/average/late) variable based on reported age at menarche (the latter to replicate Sequeira et al. [55]). The grouping into 'early' ( $\leq 1$  SD below the mean), 'average' ( $> 1$  SD below the mean and  $< 1$  SD above the mean) and 'late' onset ( $\geq 1$  SD above the mean) will be based on the distribution of age at menarche in our data. Values will be imputed for those who have not yet reached menarche at age 14, using information about the stage of pubertal development, as well as all the covariates and outcomes (see Additional file 2 for further details about the multiple imputation). We will also include self-reported breast stage at age 14 as an additional exposure for sensitivity analyses, using a scale from 0-4, from 'not yet started' to 'already complete'.

*Mental health problems.* Depressive symptoms will be assessed through the Short Mood and Feelings Questionnaire (SMFQ; 13 items) [62]. The SMFQ has demonstrated validity in general population samples of children and adolescents [63]. Anxiety symptoms will be assessed through a short form of the Screen for Child Anxiety Related Disorders (SCARED; 5 items) [64]. Behaviour problems (CD, ODD, and ADHD) will be assessed with the Rating Scale for Disruptive Behaviour Disorders (RS-DBD; 34 items) [65]. The measures will be treated as continuous, and scores will be standardised to have a mean of 0 and standard

deviation of 1. Information about psychometric properties of the scales based on the 8-year data is provided in Additional files 2 and 3. An overview of all variables included in the study, including information about informants and variable processing, is in Additional file 4.

To facilitate replication of Sequeira et al. [55], the SMFQ will also be dichotomised (with high depressive symptoms defined as scoring 11 or above). At this cut-off, the SMFQ has been found to have good sensitivity and specificity in predicting International Classification of Diseases (ICD-10) diagnosis of depression in late adolescence [66]. If the proportion of 'cases' in MoBa at 14 years using this cut-off differs from ALSPAC (based on a proportion test, comparing the percentage of cases in MoBa to the 15.5% cases in Sequeira et al.), we will run additional analyses with an adjusted cut-off defining 15.5% of the sample as cases.

*Psychiatric diagnoses.* We will link to the Control and payment of health refunds (KUHR) and the Norwegian Patient Registry (NPR) to obtain psychiatric diagnoses from medical records. KUHR covers primary health care (using codes from The International Classification of Primary Care; ICPC-2), whereas NPR covers all public specialist health-care services in Norway (using codes from ICD-10). We will extract information on diagnoses of depressive disorders (ICPC-2: P76; ICD-10: F32-F33, F34.1), anxiety disorders (ICPC-2: P74, P79, P82; ICD-10: F40-F44, F93.0-F93.2), ADHD (ICPC-2: P81; ICD-10: F90) and conduct disorders (including both CD and ODD; ICPC-2: P23; ICD-10: F91-F92). Individuals will be classified as a "case" in the case-control analysis if they have received a relevant diagnosis in either primary or secondary health care during adolescence (between age 10-17). Diagnostic status will be imputed for individuals where the beginning or the end of follow-up through the registries leads to censoring (see Additional file 2 for further details).

*Covariates.* We will include BMI at ages 8 and 14, child age at questionnaire completion, maternal and paternal age, parental education and income, financial problems, parental cohabitation, number of children in household, maternal prenatal and postnatal depression

as covariates (see Additional file 4 for further details). For replication purposes, these covariates were selected to match Sequeira et al. (see Additional file 5 for an overview).

*Genotyping and quality control.* In MoBa, blood samples were obtained from children (umbilical cord) at birth. Approximately 83,500 children have been genotyped (see Additional files 2 and 6 for further information about the genotyping) [67]. Quality control (QC) will be carried out in PLINK 1.9 and KING 2.2.5 based on the Picopili pipeline for family-based data and best-practice QC protocols in human genetics [68]. Pre-imputation QC exclusion criteria for SNPs will be: 1) known badly performing SNPs for specific genotyping arrays, 2) low minor allele frequency, 3) low genotyping call rate, 4) extreme deviation from Hardy-Weinberg equilibrium, 5) discordance in duplicate pairs of individuals, 6) association with genotype plate and genotype batch, 7) strand ambiguous (A / T and C / G), 8) not present in reference panel, or 9) different alleles than reference panel. Pre-imputation QC for individuals will be performed by filtering for: 1) heterozygosity outliers, 2) erroneous sex assignment, 3) known relatedness errors, 4) cryptic relatedness, 5) identity-by-descent, and 6) core population outliers both with and without reference to 1000 Genomes. Families with more than 5% and SNPs with more than 1% Mendelian errors will also be removed. Phasing will be performed using SHAPEIT2 with the duoHMM algorithm to incorporate pedigree information into the haplotype estimates. Imputation will be conducted using IMPUTE4. The publicly available Haplotype Reference Consortium data will be used as a reference during both phasing and imputation. We will also conduct post-imputation QC following the steps outlined in the pre-imputation QC after converting dosage data to best-guess, hard call genotype data with an imputation quality score (INFO) of 0.8 and certainty of 0.7.

*Genetic instruments for Mendelian randomisation.* A recent genome-wide association study (GWAS) meta-analysis of 42 studies involving 329,345 post-pubertal women of European ancestry found 389 independent signals associated with self-reported age at menarche, reaching the conventional threshold for genome-wide significance ( $P < 5 \times 10^{-8}$ ) in the

discovery sample [51]. These variants were largely replicated in a sample of 39,543 post-pubertal women from the Icelandic deCODE study, explaining 7.4% of the variance in age at menarche. We will first subset these genome-wide significant variants to single nucleotide polymorphisms (SNPs) only by removing insertions and deletions. Then we will extract these SNPs (as available) from the genetic data in MoBa, which did not contribute to the GWAS meta-analysis. Having subset to genome-wide significant SNPs available in the MoBa cohort, we will then clump them for independence (linkage disequilibrium  $R^2 = 0.001$ , clumping window = 10000 kb) in *MR-Base* using the *TwoSampleMR* [69] package. For the one-sample MR, we will use this set of SNPs to construct a weighted genetic risk score based on published GWAS effect estimates. The score will be computed as the weighted sum of the age-at-menarche-increasing alleles across the selected SNPs. Specifically, we will multiply the number of effect alleles (0, 1, or 2; or if imputed, probabilities of effect alleles) at each SNP by their weight (GWAS SNP-trait association), then sum and divide by the total number of SNPs used. We will also employ Steiger filtering [70] to create another genetic instrument for age at menarche, excluding SNPs that are more predictive of depression at age 14 than age at menarche. This serves primarily to prevent reverse causation, and to remove potential pleiotropic pathways other than the causal pathway of interest. The first 20 principal components will be included in all one-sample MR analyses to control for confounding by population stratification.

For the two-sample MR, the aforementioned set of SNPs will be harmonised using *MR-Base* [69], and we will infer the forward strand alleles using allele frequency information for palindromic SNPs (SNPs with minor allele frequency > 0.3 will be discarded, as these cannot be reliably inferred). For two-sample multivariable MR (MVMR) analyses accounting for estradiol levels, we will use a recently published GWAS in UK Biobank ( $N = 163,985$  females of European ancestry) which identified 4 SNPs independently associated with estradiol (at  $P < 1 \times 10^{-7}$ ) [71]. We will also conduct two-sample MR analyses to account for

potential confounding by BMI, for which the most recent female-only GWAS summary data will be used. This is due to the important assumption in two-sample MR that the employed samples stem from the same underlying population. Because of sample overlap between MoBa and the most recent GWAS of BMI in children (which identified 15 SNPs associated with childhood BMI [72]), we will use the GWAS of recalled body size at age 10 from the UK Biobank ( $N = 246,511$  females), which identified 135 SNPs independently associated with comparative early life body size [73]. For adult BMI, we will also use GWAS summary data from the UK Biobank ( $N = 246,511$  females), which identified 215 SNPs associated with adult-measured BMI [73]. The use of these measures as indicators of separate exposures has previously been validated in ALSPAC and employed in an MVMR setting [73]. Finally, to mirror how co-occurring depression is accounted for in the observational analyses of other mental health domains, we will run MVMR including a genetic instrument for depression. This will be based on the latest GWAS meta-analysis of major depressive disorder ( $N = 1,154,267$ ), which identified 223 variants independently associated with depression [74]. Note that if larger GWAS studies of European ancestry are published after the stage 1 submission, updated summary statistics will be used, and this change will be reported in the stage 2 submission.

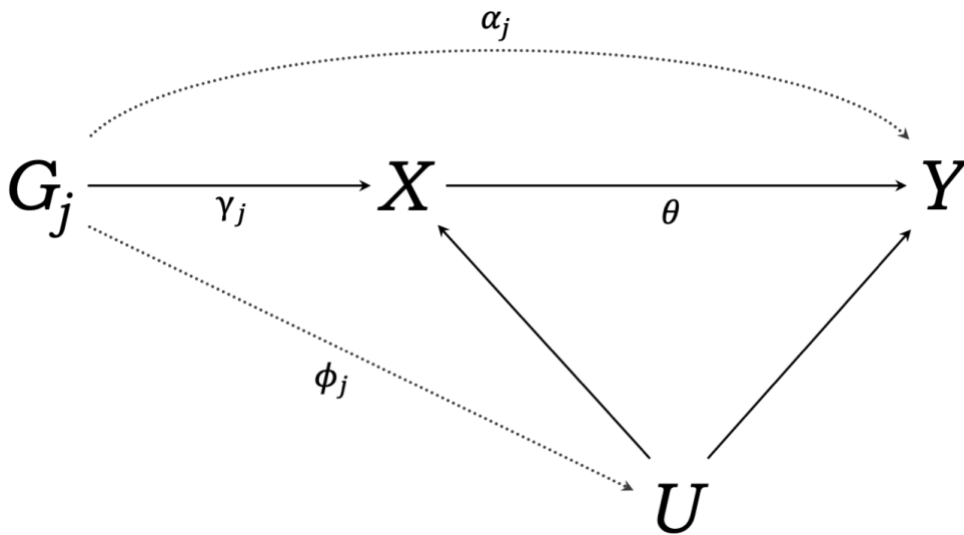
## **Statistical analysis**

### ***Observational analyses***

First, we will run linear regression analyses to estimate the observational associations between age at menarche and continuous symptom outcomes, accounting for the effects of covariates (described above). In addition, we will run logistic regression analyses to estimate observational associations with the dichotomised SMFQ and diagnostic outcomes from registry data, accounting for the effects of covariates.

## Mendelian randomisation analyses

To avoid problems related to confounding and reverse causation common to traditional observational methods, MR uses  $j$  genetic variants  $G_1, G_2, \dots, G_j$  as a proxy for the exposure  $X$  to estimate the association between the exposure  $X$  and the outcome  $Y$  (see Figure 1 for an illustrative diagram) [50]. The obtained estimate is assumed to be independent of potential confounders  $U$ . This assumption builds on Mendel's first and second law of inheritance [52]. That is, genetic variants will not generally be associated with confounders if a) the likelihood that a germ cell with a particular genetic variant contributes to a viable pregnancy is independent of the environment (Mendel's first law), and if b) genetic variants segregate independently (Mendel's second law).



**Figure 1. Directed acyclic graph illustrating the Mendelian randomisation design.**

$G_j$  is the  $j$ -th genetic variant, with effect  $\phi_j$  on confounders  $U$ , direct effect  $\gamma_j$  on exposure  $X$ , and direct effect  $\alpha_j$  on outcome  $Y$ .  $\theta$  is the estimated causal effect of the exposure on the outcome. Dotted lines represent possible violations of the MR assumptions.

*MR assumptions.* The three main assumptions of MR are: 1) that the instrument  $G_j$  is associated with the exposure  $X$ , called the relevance assumption, 2) that there are no unmeasured confounders of the gene-outcome association  $U$ , called the independence assumption, and 3) that the genetic variants  $G_j$  affect the outcome  $Y$  only through the

exposure  $X$ , called exclusion restriction. While assumption 1) can be verified empirically, assumptions 2) and 3) are empirically unverifiable (but potentially falsifiable). Owing to how instrumental variable analyses are estimated, violations of these assumptions may lead to strong biases in the estimates; therefore, such estimates should be interpreted with care and in conjunction with other evidence [75]. Several sensitivity analyses that have been developed to address potential bias from violations of the MR assumptions will be employed here.

*One-sample MR.* In the one-sample MR analyses of continuous symptom variables, we will employ two-stage least squares (2SLS) regression. In the 2SLS approach, self-reported age at menarche  $X$  is first regressed on the genetic variants  $G_j$ , obtaining the predicted values. In the second stage the regression of the outcome  $Y$  on the exposure  $X$  is estimated as usual, replacing self-reported age at menarche with the predicted values from the first stage; hereafter referred to as “genetically predicted age at menarche”. We include covariates in these analyses to increase statistical efficiency, and to control for any residual population stratification. When outcomes are excessively skewed (based on the skewness test implemented in the *moments* package in R [76]) or for binary outcomes (for which a logistic model will be used in the second stage), we will apply a post-estimation correction of the standard errors (the *HC1* option in the *sandwich* R package [77]). This replicates the *ivreg2* command in Stata with the *robust* option, used in Sequeira et al.

As a sensitivity analysis, we will carry out linear regression analyses between the genetic risk score and the covariates used for the one-sample MR. The measured covariates serve as proxies for some of the potential confounders of the relationship between the exposure and outcome. If the genetic risk score is associated with a confounder, the second assumption of MR is violated. If the risk score is associated with a potential confounder, and a suitable GWAS exists (i.e., for BMI [73], or education [78]) then we will run additional MVMR analyses with genetic instruments for those covariates included as an additional

exposure. Whether we run these follow-up analyses is contingent on the genetic instrument(s) satisfying the relevance assumption of MR (i.e., conditional instrument strength  $F > 10$ ), based on the modified  $Q_x$  statistic (see below).

We will also carry out linear regression analysis of the exposure on the genetic instrument for age at menarche and calculate the  $F$ -statistic to evaluate the relevance assumption. A common rule of thumb is that if  $F < 10$ , the instrument is considered weak [54]. Here, the criterion of  $F > 10$  for age at menarche must be met to successfully test the hypotheses that are based on one-sample MR (H3a, H3b, H4.1-4a, and H4.1-3b). In Sequeira et al. this value was  $F = 114.9$ , indicating a very strong instrument (based on a previous GWAS of age at menarche).

Combining one-sample MR with two-sample MR analyses is beneficial even when the same outcome sample is used, since any bias from weak instruments would skew the one-sample estimate towards the (confounded) observational estimate and the two-sample estimate towards the null [79]. Also, two-sample MR allows for additional sensitivity analyses that assume independence between the gene-exposure and gene-outcome estimates.

*Two-sample MR.* In the two-sample MR analyses, only the genotype-outcome ( $G$ - $Y$ ) association is estimated in MoBa. For these analyses (using the *TwoSampleMR* package [69]), we will extract estimates for the genotype-exposure ( $G$ - $X$ ) association from summary-level data from the age-at-menarche GWAS [51] and produce a set of SNP-specific Wald estimates by calculating the ratio between the  $G$ - $X$  and the  $G$ - $Y$  (estimated in MoBa) associations. We will test the heterogeneity between the Wald ratios using SNP-specific Cochran's  $Q$  statistics. Then, these estimates are combined using the inverse variance weighted (IVW) meta-analysis approach to obtain an estimate of the causal effect. We will then use several sensitivity methods to assess whether this estimate is affected by horizontal pleiotropy, further described below.



We expect pleiotropic effects in our study, which means that the employed genetic variants may affect multiple biological pathways. Specifically, pleiotropic effects via BMI are likely. Vertical pleiotropy, where the genetic variants affect other traits through the exposure, is consistent with a causal interpretation [80]. On the other hand, horizontal pleiotropic effects could bias our findings, as it violates the instrumental variable assumptions. Horizontal pleiotropy means that the genetic variants affect the outcome through other pathways than the exposure of interest.

We will employ a standard battery of two-sample sensitivity analyses, which have been developed mainly to address the third MR assumption: that the SNPs used to instrument the exposure only influence the outcome through the exposure. The standard IVW method assumes that there is no directional horizontal pleiotropy (where pleiotropic effects are biasing the estimate in the same direction), which is an assumption that is likely to be violated. Mendelian randomisation-Egger (MR-Egger) regression allows all instruments to be subject to horizontal pleiotropy but has the lowest power to detect a causal effect among these approaches [80]. The MR-Egger intercept can be used to assess bias from directional horizontal pleiotropy. MR pleiotropy residual sum and outlier (MR-PRESSO) is also used to test for the presence of directional pleiotropy, and additionally detect outliers [81]. The MR-PRESSO test consists of three parts: a) the global test, which detects horizontal pleiotropy, b) the outlier-corrected causal estimate, which corrects for any detected pleiotropy, and c) the distortion test, which tests whether the MR estimate is significantly different after adjustment for outliers. In addition, we will use the weighted median estimator, which assumes that the genetic variants representing over 50% of the weight in the analysis are valid instruments [82]. The contamination mixture method assumes that only some of the genetic variants are valid instruments and has a well-controlled type 1 error rate compared to similar approaches [83]. These five methods make different (to some extent orthogonal)

assumptions about horizontal pleiotropy, and we consider effects that are consistent across these approaches to be more likely causal, in line with a triangulation approach.

*Multivariable MR.* It is also possible to estimate the effects of multiple exposures simultaneously, using MVMR [84]. Including multiple genetic instruments in the same analysis is valuable when the exposures being studied are highly correlated. Here, we will use summary data (described above) and the *MVMR* package (<https://github.com/WSpiller/MVMR/>) to estimate the direct causal effect of age at menarche on mental health while adjusting for BMI. Within the MVMR setting, the first MR assumption is that the instrument is robustly associated with the exposure, conditional on the remaining exposures in the model [85]. This is quantified by the modified  $Q_x$  statistic (estimating heterogeneity in gene-exposure effects), where a higher degree of heterogeneity indicates greater instrument strength. We will convert  $Q_x$  to the conventional  $F$  statistic to evaluate conditional instrument strength (with a threshold of  $F > 10$  for each instrument). The second and third assumptions in MVMR are direct extensions of the univariate MR assumptions described above. We will employ MVMR-Egger, MVMR-Median, and MVMR-Lasso [86] as sensitivity analyses to test for violations of these assumptions in the multivariable setting.

We will also run two-sample MVMR with estradiol included as an exposure, although this analysis will likely be limited by imbalanced instruments. If  $F < 10$  based on the modified  $Q_x$  statistic, we will restrict our interpretation to the direct effect of age at menarche accounting for estradiol and avoid interpreting the effect of estradiol on the outcome. For this analysis, leave-one-out analyses and single-SNP plots will be used as sensitivity analyses due to the limited number of SNPs available.

### ***Equivalence testing***

Because null hypothesis significance testing cannot support substantive interpretations of non-significant results, we will employ equivalence testing to quantify support for the null

hypothesis [87]. This is done by testing whether the 90% confidence intervals (CIs) for the effect size overlap with pre-specified equivalence bounds for the smallest effect size of interest (SESOI). Note that we use the 90% CIs for the observational analyses rather than the 95% CIs, as the effect size is tested against two equivalence bounds (the lower and upper) with two separate one-tailed tests. If the 90% CIs are inside the equivalence bounds, this will indicate that the effect size is surprisingly small given that an effect as large as the SESOI exists. If the 90% CIs are not entirely within the equivalence bounds, this will indicate that the effect size is of a meaningful magnitude.

### ***Negative control analyses***

Since an individual's age at menarche cannot influence their mental health prior to puberty, childhood symptoms can serve as a negative control outcome in our study. Such analyses can be used to detect unmeasured confounding, also in the context of MR, given that the negative control outcome is associated with confounders in a similar way to the outcome of interest [88]. Here, we will focus on comparing the estimates for continuous outcomes before puberty (at 8 years) and after puberty (at 14 years), by employing equivalence testing to determine whether the 14-year estimate is statistically equivalent to zero after accounting for unmeasured confounding by setting the equivalence bound to the upper bound of the 8-year estimate.

## **Sampling plan**

### ***Inclusion criteria and sample size***

We will include all MoBa females (as registered at birth in the Medical Birth Registry; MBRN) with any available phenotype data. We have a sample size of  $N = 20,225$  females with phenotype data and expect  $N \simeq 15,500$  (genotype QC is ongoing) with both phenotype and genotype data at age 8. The observed response rate has been between 30-34% in those

who have already received the 14-year MoBa questionnaire. Conservatively projecting a 30% return rate, we expect  $N \simeq 12,500$  females with phenotype data and  $N \simeq 10,000$  with both phenotype and genotype data at age 14. We will run the observational analyses in the largest available sample of 14-year questionnaire responders, and then restrict these analyses to only genotyped individuals as a sensitivity analysis.

We have estimated the expected prevalence of diagnostic outcomes, derived from registry data, at the anticipated time of analysis and in the analytic sample (14-year questionnaire returners with and without genetic data). To do this we calculated rates for all relevant diagnoses among female MoBa participants who already have registry follow-up for all years between the ages of 10 and 17 (inclusive). We projected these onto 100 random samples (50 with  $N = 12,500$  and 50 with  $N = 10,000$ ) of girls whose mothers responded to the previous MoBa questionnaire (see Additional file 7: Supplementary Code for full details). The projected prevalence rates were 13.5% for depression, 16.1% for anxiety, 2.5% for CD/ODD, and 13.9% for ADHD.

*Missingness/handling of missing data.* Within the 14-year sample, we will use multiple imputation (MI) to account for missing data, specifically because of two anticipated forms of censoring in the data. The first is that some individuals will likely report having not yet had their first menstrual period in the 14-year questionnaire. Imputed values for age at menarche for these individuals will not be allowed to be lower than 14 years. The second form of censoring is in the linked registry data, which has missing data about early life diagnoses for the oldest MoBa participants (because linkage is only available since 2008). In addition, at the time of carrying out the analyses, the linked registries will have missing data about diagnoses in the later years of adolescence for younger MoBa participants (because they will not have yet turned 18 by 2021, the last year for which complete registry linkage is expected to be available). Further details on the multiple imputation are presented in

Additional file 2. In addition to MI, we will use inverse probability weighting to address potential bias from selective attrition out of the study over time (details in Additional file 2).

*Definition/removal of outliers.* We will retain all valid responses to the questionnaires. If participants ticked multiple boxes on a single-response item, their response on this item is set to missing. If respondents complete less than half of the items for a given scale, their scale score is not computed and their data for this variable is considered missing. For other phenotype data, values > 4 standard deviations from the mean will be treated as outliers and coded as missing (e.g., to remove implausible height/weight values used to calculate BMI).

### ***Power calculations***

Power analyses were conducted in R by simulation for all null hypothesis significance tests (NHSTs) and equivalence tests used to investigate each hypothesis. Details of the simulations and analyses are given in brief below, with full information and code presented in Additional file 7.

*Data generation.* For all power analyses, we simulated 1000 replicate datasets under various possible scenarios of data availability and effect sizes. For analyses with diagnoses as outcomes, we also varied case rates within plausible ranges. For all replicates across all scenarios and analyses, we began by simulating an age at menarche variable, sampling from a normal distribution with a mean of 151.52 (months) and standard deviation of 14.11 (values from Sequeira et al. [55]), rounded to the nearest year (to replicate the response format of the age at menarche item in the MoBa 14-year questionnaire). The simulated variable was allowed to take values > 14 years, meaning that all power analyses assume imputation of this variable (see Additional file 2) has already taken place. Details of the other variables simulated in the power analyses for each hypothesis are given below.

*Power calculation.* Power for NHSTs (detailed below) was calculated empirically as the proportion of replicates in each scenario in which the null hypothesis was rejected with a 5% alpha. Power for equivalence tests (detailed in Additional file 2) was calculated empirically as the proportion of replicates in each scenario in which the null hypothesis of the equivalence test (i.e., in the case of a two-tailed test, that an effect is not equivalent to zero; and in the case of a one-tailed test, that an effect is not smaller than the SESOI) was rejected with a 5% alpha.

*Hypotheses 1a/b.* For H1a we additionally simulated a depressive symptoms variable ( $M = 5.71$ ,  $SD = 4.93$ ; values from Sequeira et al. [55]), with a floor effect at 0. In the range of scenarios simulated for this hypothesis, the correlation between age at menarche and depressive symptoms was specified at each  $r$  from 0 to -0.15 in increments of 0.01, for sample sizes of 12,000 and 13,000 respectively. Results indicated 95% power to detect an effect of Cohen's  $D \geq 0.08$  at  $N = 12,000$  and  $\geq 0.06$  at  $N = 13,000$  (full results for all scenarios are presented in Additional file 8: Figure S1). For H1b we additionally simulated a binary depression diagnosis variable, based on a pre-specified prevalence and association with age at menarche (varied across scenarios at, respectively, 2%, 6%, 10%, 14% and  $D = 0$  to -0.26 in increments of 0.02). Results indicated 95% power to detect an effect of Cohen's  $D \geq 0.14$  at both  $N = 12,000$  and  $N = 13,000$  for depression prevalence of 6% or higher (full results for all scenarios are presented in Additional file 8: Figure S2).

*Hypotheses 2.1-4a/2.1-3b.* For H2.1-4a we additionally simulated anxiety, CD, ODD, and ADHD symptoms, with distributions based on the same variables in MoBa data at 8 years. As for hypothesis 1a, we simulated data for scenarios with age at menarche-outcome correlations specified from 0 to -0.15 in increments of 0.01, for sample sizes of 12,000 and 13,000 respectively. Results indicated 95% power to detect an effect of Cohen's  $D \geq 0.12$  at both simulated sample sizes (full results for all scenarios are presented in Additional file 8: Figure S3). For H2.1-3b, the simulation was essentially identical to H1b (with different

equivalence bounds and use of two-tailed tests - see *Analysis plan* below for details).

Results indicated 95% power to detect an effect of Cohen's  $D \geq 0.20$  at  $N = 12,000$  and  $\geq 0.18$  at  $N = 13,000$  for diagnosis prevalence of 2% or higher (full results for all scenarios are presented in Additional file 8: Figure S4).

*Hypotheses 3a/b.* For H3a we additionally simulated a genetic instrument for age at menarche ( $M = 0$ ,  $SD = 1$ ), and a depressive symptoms variable as per H1a. The  $R^2$  of the genetic instrument for the simulated age at menarche variable was set at 0.05, 0.075, and 0.10 across scenarios. The association between simulated age at menarche and the depressive symptoms variable was based on the average causal effect (specified from  $D = 0$  to -0.30 in increments of -0.01 across scenarios) plus an observational confounding effect (drawn randomly from a normal distribution  $M = 0$ ,  $SD = 0.05$  on the  $D$  scale for each replicate). Results indicated 95% power to detect an average causal effect (using 2SLS MR) of Cohen's  $D \geq 0.2$  when the  $R^2$  of the instrument is 0.075 or above at either simulated sample size (full results for all scenarios are presented in Additional file 8: Figure S5). For H3b we again simulated the genetic instrument for age at menarche at  $R^2$  0.05, 0.075, and 0.10 across scenarios. As per H1b, a depressive diagnosis outcome was simulated with prevalence ranging from 0.02 to 0.14 across scenarios, with its relationship to simulated age at menarche parameterised as the causal odds (specified from  $D = 0$  to -0.30 in increments of -0.02 across scenarios) plus an observational confounding effect as in H3a above. Results indicated 95% power to detect a causal effect (using logistic 2SLS MR with robust standard errors) of Cohen's  $D \geq 0.24$  when the  $R^2$  of the instrument is  $\geq 0.075$  and depression prevalence 14% at either simulated sample size (full results for all scenarios are presented in Additional file 8: Figure S6).

*Hypotheses 4.1-4a/4.1-3b.* Simulations for H4.1-4a were essentially identical to those for H3a (with different equivalence bounds and use of two-tailed tests - see *Analysis plan* below for details). Results indicated 95% power to detect an average causal effect (using 2SLS

MR) of Cohen's  $D \geq 0.2$  when the  $R^2$  of the instrument is 0.10 at either simulated sample size and 80% power to detect Cohen's  $D \geq 0.2$  with the  $R^2$  of the instrument at  $\geq 0.075$  (full results for all scenarios are presented in Additional file 8: Figure S7). Simulations for H4.1-3b were essentially identical to those for H3b (with different equivalence bounds and use of two-tailed tests - see *Analysis plan* below for details). Results indicated 95% power to detect a causal effect (using logistic 2SLS MR with robust standard errors) of Cohen's  $D \geq 0.26$  when the  $R^2$  of the instrument is  $\geq 0.075$  and diagnosis prevalence is 14% in either simulated sample size and 80% power for Cohen's  $D \geq 0.20$  in the same scenarios (full results for all scenarios are presented in Additional file 8: Figure S8).

## **Analysis plan**

We will conduct all statistical analyses in R version  $> 4$ . Below we describe the statistical analyses that will be used to test each of the hypotheses. Each null hypothesis significance test performed as part of these analyses corresponds to a single theoretical prediction. This allows alpha for each test to be preserved at 5% without multiple testing correction, but substantially narrows the nature of the theoretical model being investigated in each case, and accordingly reduces the breadth of the conclusions that can be drawn. Statistical tests and inference criteria for each hypothesis are detailed further in Additional file 1. The MR results will be reported according to Strengthening the reporting of observational studies in epidemiology using Mendelian randomisation (STROBE-MR) guidelines [89, 90]. All models will be run with and without the covariates described above to obtain adjusted and unadjusted estimates, and inferences will be based on the adjusted estimates. All models will also be run with both a continuous and categorical score for self-reported age at menarche, to facilitate replication of Sequeira et al. [55] where a categorised score was used. Inferences from the main analyses will be based on the estimates using the continuous age at menarche score.



**Question 1: To what extent is age at menarche associated with adolescent depression?**

*Hypothesis 1a.* First, we hypothesise that **earlier age at menarche will be associated with elevated depressive symptoms at age 14**. We will run a linear regression model with age of menarche as the independent variable and depressive symptoms at age 14 as the dependent variable. We will then add 8-year depressive symptoms as a covariate, to examine whether age at menarche is associated with post-pubertal symptoms independent of pre-pubertal symptoms. We will also run the same analyses with a dichotomised version of the SMFQ as the outcome, as in Sequeira et al.

In line with our directional hypothesis and one-tailed null hypothesis significance test, we will apply an equivalence test only using an upper bound (sometimes called an “inferiority” test). This tests whether the null hypothesis of an effect at least as large as the SESOI can be rejected. We derive the SESOI for this analysis based on the lower end of the confidence interval of a meta-analytic estimate (as recommended by Lakens, Scheel, and Isager [91]) of age at menarche and depressive symptoms in adolescents. This conservative approach was deemed appropriate given the potential for bias in the literature (estimated to have a small and positive impact on the magnitude of results for early pubertal timing and internalising behaviours in the meta-analysis by Ullsperger and Nikolas [20]). Because they did not estimate the specific association with depression, we conducted a meta-analysis of the studies of community adolescents that they included [24, 29, 31, 38, 49] which had data on age of menarche and depressive symptoms (see Additional file 7 for further details, and the script used to run the meta-analysis). The pooled association of 5 eligible studies of age at menarche and depressive symptoms in early-to-mid adolescence was  $D = 0.28$  (95% CI = 0.23 - 0.33). Our SESOI is the lower CI bound of this estimate (i.e.,  $D = 0.23$ ).

We will make the inference that hypothesis 1a is supported if 1) the coefficient for the effect of age at menarche on 14-year depressive symptoms is significantly less than zero (one-tailed test; alpha 5%) in the pre-pubertal symptoms-adjusted model; and 2) we fail to reject the null hypothesis that this effect in the population is at least as large as the SESOI (one-tailed test; alpha 5%). The interpretation of all potential patterns of results is in Additional file 1.

*Sensitivity analyses.* In addition, we will run a linear regression with breast stage instead of age at menarche included as the predictor, and then a multiple regression with both breast stage and age at menarche included as predictors, replicating previous observational analyses of breast stage in ALSPAC [42]. Both models will be adjusted for 8-year depressive symptoms. Breast stage is an indicator of pubertal stage, which at any point in time during adolescence will be more advanced in those who began puberty earlier (i.e., those with earlier age at menarche). Thus, we first investigate whether breast stage is associated with 14-year depressive symptoms in isolation, and then the relative contribution of each in a multivariable model including both breast stage and age at menarche. This sensitivity analysis will be repeated for all subsequent observational analyses.

*Hypothesis 1b.* We hypothesise that **earlier age at menarche will be associated with higher rates of depression diagnoses during adolescence**. We will run a logistic regression model with age of menarche as the independent variable and depression diagnosis as the dependent variable. This analysis will include any depression diagnosis in either primary or secondary health care during adolescence (age 10-17). We will additionally co-vary for depression status prior to puberty (age 0-8) if rates of pre-puberty diagnoses are sufficiently high and evenly distributed across levels of the outcome variable to allow model convergence.

We will make the inference that hypothesis 1b is supported if 1) the coefficient for the effect of age at menarche on odds of depression diagnoses is significantly less than zero (one-tailed test; alpha 5%) in the pre-pubertal depression status-adjusted model; and 2) we fail to reject the null hypothesis that this effect in the population is at least as large as the SESOI (one-tailed test; alpha 5%).

***Question 2: Does age at menarche associate with symptoms or diagnoses in other domains of mental health, independent of depression?***

*Hypotheses 2.1-4a.* We hypothesise that **the association with age at menarche will extend to other symptom domains: anxiety (H2.1a); CD (H2.2a); ODD (H2.3a); and ADHD (H2.4a)**. To test each hypothesis, we will run linear regression models to examine associations between age at menarche and each symptom domain at age 14. We will then add depressive symptoms at age 14 as a covariate in each model to examine whether any associations in other domains are independent of co-occurring depressive symptoms. Finally, we will also add a measure of each symptom domain at age 8 as a covariate in the age 14 model for that domain, to examine whether associations between age at menarche and post-pubertal symptoms are additionally independent of pre-pubertal symptoms (these are referred to as the “fully adjusted” models below).

We will use the lower end of the CIs of a meta-analytic estimate of age at menarche and general psychopathology in adolescents to determine the SESOI across domains because precise meta-analytic estimates are not available or feasible to derive for each of the domains (e.g., the number of studies of age at menarche and ADHD is limited). The pooled association of 42 studies in Ullsperger and Nikolas 20 was  $D = 0.27$  (95% CI = 0.22 - 0.31). Based on the lower end of the CIs we will use equivalence bounds of -0.22 - 0.22 in our analysis.

We will infer support for hypotheses 2.1-4a if 1) the coefficient for the association between age at menarche and a domain of 14-year symptoms in the fully adjusted model is different from zero (two-tailed tests, 5% alpha); and 2) we fail to reject the null hypothesis that the association in the population is at least as extreme as the SESOI in either direction (two one-tailed tests, 5% alpha). The interpretation of all potential patterns of results is in Additional file 1.

*Hypotheses 2.1-3b.* We hypothesise that **the association with earlier age at menarche will extend to diagnoses in other mental health domains: anxiety disorders (H2.1b); conduct disorders (H2.2b), including CD and ODD; and ADHD (H2.3b).** For these analyses, we will run logistic regression models to examine associations between age at menarche and odds of receiving each diagnosis during adolescence (age 10-17). We will first add depression diagnostic status (age 10-17) as a covariate, then pre-pubertal diagnostic status (age 0-8) in each relevant domain to each model, to examine whether associations between age at menarche and diagnoses in other domains are independent of both comorbid depression and prior diagnoses.

We will infer support for hypotheses 2.1-3b if 1) the coefficient for the association between age at menarche and a diagnosis in the fully adjusted model is different from zero (two-tailed tests, 5% alpha); and 2) we fail to reject the null hypothesis that the association in the population is at least as extreme as the SESOI in either direction (two one-tailed tests, 5% alpha).

***Question 3: What is the evidence for a causal link between age at menarche and depression?***

*Hypothesis 3a.* We hypothesise that **earlier age at menarche will show a causal relationship with elevated depressive symptoms at age 14.** In the one-sample MR analysis, we will use 2SLS regression to test the relationship between the genetic risk score

for age at menarche and depressive symptoms at age 14 (see Methods for a further description). We will also run the same analyses with a dichotomised version of the SMFQ as the outcome, answering whether we can replicate the result at 14 years in Sequeira et al.

To test for potential bias in this estimate from unmeasured confounding, we will conduct a negative control MR analysis using depressive symptoms prior to puberty as a negative control outcome. Specifically, we will run the same 2SLS model with depressive symptoms at 8 years as the outcome. A relationship between genetically predicted age at menarche and childhood depressive symptoms would be temporally implausible, indicating unmeasured confounding. There are no established statistical procedures to refine the MR estimate by factoring in the negative control outcome. It has been suggested that calibrating the putative causal estimate with a quantitative contrast between the negative control and the main estimate could lead to bias (for further detail see Sanderson et al. [88]). Therefore, we will focus on testing for the degree of confounding rather than refining the MR estimate. To formally test whether the extent of observed confounding is sufficient to account for the observed effect at 14-years, we will subject the 14-year effect to an equivalence test, setting the equivalence bound to the upper bound of the 8-year estimate.

We used the “small telescopes” approach [92] for setting the SESOI, which is particularly suitable for replications. In this approach, the SESOI is set to the effect size that the original study had 33% power to detect. The idea is that based on this power level, the probability of observing an effect (if a true effect exists) is too low to reliably distinguish signal from noise. We calculated the effect size the original study would have 33% power to detect using the *mRnd* power calculator for Mendelian randomisation [93]. We used the values from the original study [55] to determine the effect size ( $N = 2,404$ ,  $\alpha = 0.05$ ,  $K = 0.155$ ,  $R^2_{XG} = 0.049$ ) where  $N$  is the sample size,  $\alpha$  is the Type-I error rate,  $K$  is the proportion of cases in the study, and  $R^2_{XG}$  is the proportion of variance explained for the association between the genetic variants  $G_j$  and the exposure  $X$ . The resulting effect size was  $D = 0.25$ , which was

selected as the SESOI for this analysis. In line with our directional hypothesis and one-tailed null hypothesis significance test, we again apply an inferiority test (setting and testing on an equivalence bound at the SESOI only in the predicted direction).

We will infer support for hypothesis 3a (that earlier age at menarche causes elevated depressive symptoms at age 14) if 1) the coefficient for the causal effect of age at menarche on 14-year depressive symptoms is significantly less than zero (one-tailed test; alpha 5%); 2) we fail to reject the null hypothesis that this causal effect in the population is at least as large as the SESOI (one-tailed test; alpha 5%); and 3) we fail to reject the null hypothesis that this causal effect in the population is at least as large as the upper bound of the negative control (8-year) estimate (one-tailed test; alpha 5%). The interpretation of all patterns of results is described in Additional file 1.

*Sensitivity analyses.* The traditional MR approach described for H3a above assumes that there is no horizontal pleiotropy. We expect that the most likely threat to this assumption is pleiotropy via childhood body size/BMI. Previous studies have attempted to solve this by excluding SNPs associated with childhood (as a proxy for ‘pre-pubertal’) and/or adult (as a proxy for ‘post-pubertal’) BMI [55]. However, excluding adult BMI SNPs in particular risks inducing a spurious association with depression, due to collider bias [94]. We will therefore conduct an MVMR analysis with genetic instruments for age at menarche, childhood body size and adult BMI included in the same model - which estimates the direct effect of age at menarche on the outcome. MVMR accounts for any overlap analytically, whereas excluding SNPs associated with BMI based on *P*-values will likely miss SNPs below the employed threshold. Finally, like the MVMR with BMI, we will also run a model including the genetic instrument for estradiol to test the direct effect of age at menarche on depressive symptoms when accounting for estradiol (see Methods).

In addition, we will conduct several sensitivity analyses to assess the three main assumptions of MR: 1) that the instrument is associated with the exposure, 2) that the genetic variants are independent of all confounders, and 3) that the instrument affects the outcome only through the exposure of interest. To evaluate the assumptions, we use 1)  $F$ -statistics for the instrument-exposure association (where  $F > 10$  is required), 2) regression of the covariates on the genetic risk score (statistically significant relationships indicate potential confounding), and 3) MR sensitivity analyses. In both the one-sample and two-sample setting we will exclude SNPs that are associated with more variation in depression than age at menarche (see Methods for a description). If the causal relationship is then attenuated, this may suggest the existence of other pleiotropic pathways or reverse causation. We will also employ two-sample MR sensitivity analyses: MR-Egger, MR-PRESSO, weighted median and the contamination mixture method (see Methods). The MR-Egger intercept and MR-PRESSO global test are used to test for bias from directional horizontal pleiotropy. For MR-Egger, a significant intercept will be considered indicative of bias from directional horizontal pleiotropy. For MR-PRESSO, we will report the outlier-corrected causal estimate if both the global test and the distortion test are significant. More broadly, the purpose of these sensitivity tests is to ensure that the MR results are valid, and violations of any MR assumptions will therefore temper our inferences. However, sensitivity analyses such as MR-Egger are subject to their own biases, and therefore the strongest indication that results are unlikely to be biased by horizontal pleiotropy would be consistent evidence across the different methods.

*Hypothesis 3b.* We further hypothesise that **earlier age at menarche will result in higher rates of depression diagnoses during adolescence**. As for hypothesis 1b, we will run these binary outcome MR analyses with individuals diagnosed either in primary or secondary health care during adolescence as cases (age 10-17). We will also run the same MR sensitivity analyses as in H3a, including the negative control MR analysis using depression

diagnoses during childhood as the outcome (age 0-8). Here, we will compare the direction, magnitude, and precision of the estimates for the main outcome and negative control outcome to determine the degree of unobserved confounding.

We will infer support for hypothesis 3b if 1) the coefficient for the causal effect of age at menarche on depressive disorders is significantly less than zero (one-tailed test; alpha 5%); 2) we fail to reject the null hypothesis that this causal effect in the population is at least as large as the SESOI (one-tailed test; alpha 5%).

*Sensitivity analyses.* Here, we will conduct all sensitivity analyses described in H3a above.

***Question 4: Is there evidence of causal links between age at menarche and other domains of mental health?***

*Hypothesis 4.1-4a.* We hypothesise that **age at menarche will be causally linked with symptoms in other domains: anxiety (H4.1a); CD (H4.2a); ODD (H4.3a); and ADHD (H4.4a)**. To test each of these hypotheses, we will use 2SLS regression to investigate the relationship between the genetic risk score for age at menarche and each symptom domain at age 14. We will also run negative control analyses using the corresponding symptom domains at age 8 as negative control outcomes. For these analyses, there were no prior MR studies to base our estimated effect size on. Therefore, our SESOI will be  $D = 0.20$  (i.e., what is considered a small effect size, in the absence of a clear theoretical justification). Thus, the equivalence bounds will be  $-0.20 - 0.20$ .

We will infer support for hypotheses 4.1-4a if 1) the coefficient for the causal effect of age at menarche on a domain of 14-year symptoms is different from zero in the fully adjusted model (two-tailed tests, 5% alpha); 2) we fail to reject the null hypothesis that the causal effect in the population is at least as extreme as the SESOI in either direction (two one-tailed tests, 5% alpha); and 3) we fail to reject the null hypothesis that the causal effect in the



population is at least as large as the upper bound of the negative control (8-year) estimate (one-tailed test; alpha 5%). The interpretation of all patterns of results is in Additional file 1.

*Sensitivity analyses.* Here, we will conduct the two-sample sensitivity analyses described in H3a above (MR-Egger, MR-PRESSO, weighted median and contamination mixture). We will also conduct MVMR analyses accounting for BMI and estradiol. In addition, we will seek to account for the potential overlap between depression and symptoms in other domains (mirroring the observational analyses) by including a genetic instrument for depression alongside age at menarche in an additional MVMR analysis (see Methods).

*Hypothesis 4.1-3b.* Here, we hypothesise that **the genetic risk score for age at menarche will be associated with diagnoses in other domains: anxiety disorders (H4.1b); conduct disorders (H4.2b), including CD and ODD; and ADHD (H4.3b).** To test each of these hypotheses, we will use 2SLS regression to test the relationship between the genetic risk score for age at menarche and diagnoses of each condition. These analyses will include any relevant diagnosis in either primary or secondary health care during adolescence (age 10-17). As for H4.1-4a, we will run negative control analyses using the corresponding conditions during childhood as outcomes (age 0-8). We will compare the direction, magnitude, and precision of the estimates for the main outcomes and negative control outcomes to determine the degree of unobserved confounding. Here, we will use equivalence bounds of Cohen's  $D$  -0.20 - 0.20, as in H4.1-4a.

We will infer support for hypotheses 4.1-3b if 1) the coefficient for the causal effect of age at menarche on a diagnosis is different from zero (two-tailed tests, 5% alpha); and 2) we fail to reject the null hypothesis that the causal effect in the population is at least as extreme as the SESOI in either direction (two one-tailed tests, 5% alpha).

*Sensitivity analyses.* Here, we will run the same sensitivity analyses as in H4.1-4a above.

## Abbreviations

**MoBa:** Norwegian Mother, Father, and Child Cohort Study

**ADHD:** Attention-deficit hyperactivity disorder

**ALSPAC:** Avon Longitudinal Study of Parents and Children

**BMI:** Body mass index

**MR:** Mendelian randomisation

**SNPs:** Single nucleotide polymorphisms

**RCT:** Randomised controlled trial

**CD:** Conduct disorder

**ODD:** Oppositional defiant disorder

**SMFQ:** Short Mood and Feelings Questionnaire

**SCARED:** Screen for Child Anxiety Related Disorders

**RS-DBD:** Rating Scale for Disruptive Behaviour Disorders

**ICD-10:** International Statistical Classification of Diseases and Related Health Problems

**KUHR:** Control and payment of health refunds

**NPR:** Norwegian Patient Registry

**ICPC-2:** The International Classification of Primary Care

**QC:** Quality control

**GWAS:** Genome-wide association study

**MVMR:** Multivariable Mendelian randomisation

**2SLS:** Two-stage least squares

**IVW:** Inverse variance weighted

**MR-Egger:** Mendelian randomisation-Egger

**MR-PRESSO:** MR pleiotropy residual sum and outlier

**MVMR-Egger:** Multivariable Mendelian randomisation-Egger

**MVMR-Median:** Multivariable Mendelian randomisation-Median

**MVMR-Lasso:** Multivariable Mendelian randomisation-Lasso

**CIs:** Confidence intervals

**SESOI:** Smallest effect size of interest

**MI:** Multiple imputation

**NHST:** Null hypothesis significance testing

**STROBE-MR:** Strengthening the reporting of observational studies in epidemiology using Mendelian randomisation

## **Declarations**

### **Ethics approval and consent to participate**

The establishment of MoBa and initial data collection was based on a license from the Norwegian Data Protection Agency and approval from The Regional Committees for Medical and Health Research Ethics. The MoBa cohort is now based on regulations related to the Norwegian Health Registry Act. The current study was approved by The Regional

Committees for Medical and Health Research Ethics (REK numbers 2016/1702). By consenting to MoBa, participants have also agreed to linkage to KUHR, NPR, and MBRN. MBRN is a national health registry containing information about all births in Norway.

## **Consent for publication**

Not applicable.

## **Availability of data and materials**

The MoBa data are not publicly available as the consent given by the participants does not open for storage of data on an individual level in repositories or journals. Researchers who want access to data sets for replication should submit an application to [datatilgang\(at\)fhi.no](mailto:datatilgang(at)fhi.no). Access to datasets requires approval from The Regional Committee for Medical and Health Research Ethics in Norway and an agreement with MoBa.

Data preparation and analysis code for all elements of the project will be made publicly available on Github at <https://github.com/psychgen/aam-psych-adolesc-rr>.

## **Competing interests**

The authors declare that they have no competing interests.

## **Funding**

The Research Council of Norway supports F.A.T., C.S., E.C., H.A., T.R.-K., N.M.D., and O.A.A. (#300668; #274611; #273659, #273659, #324620, #274611, #295989, #229129; #213837; #248778; #223273; #249711). The South-Eastern Regional Health Authority supports A.D.A., R.E.W., O.A.A., A.H., and L.J.H. (#2020023, #2020024, 2017-112, #2020022, #2018058). The National Institute of Mental Health supports R.B.L. (#R01-MH101269). N.M.D. and G.D.S. work in a unit that receives support from the University of

Bristol and the UK Medical Research Council (MC\_UU\_00011/1). O.A.A. is also supported by Stiftelsen Kristian Gerhard Jebsen and H2020 grant CoMorMent (#847776). This work was partly supported by the Research Council of Norway through its Centres of Excellence funding scheme (#262700). The funders have/had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

## **Author's contributions**

Author contributions are presented according to the CRediT (Contributor Roles Taxonomy).

A.D.A: Conceptualisation, methodology, formal analysis, software, visualization, writing - original draft, writing - review & editing, project administration. R.E.W: Conceptualisation, methodology, software, writing - review & editing. F.A.T: Writing - review & editing. R.B.L: Writing - review & editing. H.A: Writing - review & editing, funding acquisition. E.C: Data curation, software, writing - review & editing. M.C.M: Conceptualisation, methodology, writing - review & editing. T.R.-K: Writing - review & editing, funding acquisition. P.M: Investigation, writing - review & editing. O.A.A: Investigation, writing - review & editing. C.S.: Writing - review & editing. G.D.S: Writing - review & editing. N.M.D: Methodology, writing - review & editing. A.H: Conceptualisation, methodology, writing - review & editing, supervision, funding acquisition. L.J.H: Conceptualisation, methodology, formal analysis, software, data curation, writing - original draft, writing - review & editing, supervision, funding acquisition. All authors read and approved the final manuscript.

## **Acknowledgements**

We thank the Norwegian Institute of Public Health (NIPH) for generating high-quality genomic data. This research is part of the HARVEST collaboration, supported by the Research Council of Norway (#229624). We also thank deCODE Genetics, and the NORMENT Centre for providing genotype data, funded by the Research Council of Norway

(#223273), South-Eastern Norway Health Authority and KG Jebsen Stiftelsen. We further thank the Center for Diabetes Research, the University of Bergen for providing genotype data and performing quality control and imputation of the data funded by the ERC AdG project SELECTIONPREDISPOSED, Stiftelsen Kristian Gerhard Jebsen, Trond Mohn Foundation, the Research Council of Norway, the Novo Nordisk Foundation, the University of Bergen, and the Western Norway health Authorities (Helse Vest). The Norwegian Mother, Father and Child Cohort Study is supported by the Norwegian Ministry of Health and Care Services and the Ministry of Education and Research. We are grateful to all the participating families in Norway who take part in this on-going cohort study.

## References

1. Benoit A, Lacourse E, Claes M. Pubertal timing and depressive symptoms in late adolescence: The moderating role of individual, peer, and parental factors. *Dev Psychopathol.* 2013;25:455–71.
2. Conley CS, Rudolph KD, Bryant FB. Explaining the longitudinal association between puberty and depression: sex differences in the mediating effects of peer stress. *Dev Psychopathol.* 2012;24:691–701.
3. Copeland W, Shanahan L, Miller S, Costello EJ, Angold A, Maughan B. Outcomes of early pubertal timing in young women: a prospective population-based study. *Am J Psychiatry.* 2010;167:1218–25.
4. Ge X, Conger RD, Elder Jr GH. Pubertal transition, stressful life events, and the emergence of gender differences in adolescent depressive symptoms. *Dev Psychol.* 2001;37:404–17.
5. Ge X, Kim IJ, Brody GH, Conger RD, Simons RL, Gibbons FX, et al. It's about timing and change: pubertal transition effects on symptoms of major depression among African American youths. *Dev Psychol.* 2003;39:430–9.
6. Ge X, Brody GH, Conger RD, Simons RL. Pubertal maturation and African American children's internalizing and externalizing symptoms. *J Youth Adolesc.* 2006;35:528–37.
7. Graber JA, Seeley JR, Brooks-Gunn J, Lewinsohn PM. Is pubertal timing associated with psychopathology in young adulthood? *J Am Acad Child Adolesc Psychiatry.* 2004;43:718–26.
8. Graber JA, Brooks-Gunn J, Warren MP. Pubertal effects on adjustment in girls: Moving from demonstrating effects to identifying pathways. *J Youth Adolesc.* 2006;35:391–401.
9. Hamlat EJ, Stange JP, Abramson LY, Alloy LB. Early pubertal timing as a vulnerability to depression symptoms: Differential effects of race and sex. *J Abnorm Child Psychol.* 2014;42:527–38.
10. Keenan K, Culbert KM, Grimm KJ, Hipwell AE, Stepp SD. Timing and tempo: Exploring the complex association between pubertal development and depression in African American and European American girls. *J Abnorm Psychol.* 2014;123:725–36.
11. Mendle J, Harden KP, Brooks-Gunn J, Graber JA. Development's tortoise and hare: pubertal timing, pubertal tempo, and depressive symptoms in boys and girls. *Dev*

- Psychol. 2010;46:1341–53.
12. Nadeem E, Graham S. Early puberty, peer victimization, and internalizing symptoms in ethnic minority adolescents. *J Early Adolesc.* 2005;25:197–222.
  13. Rudolph KD, Troop-Gordon W. Personal-accentuation and contextual-amplification models of pubertal timing: Predicting youth depression. *Dev Psychopathol.* 2010;22:433–51.
  14. Blumenthal H, Leen-Feldner EW, Babson KA, Gahr JL, Trainor CD, Frala JL. Elevated social anxiety among early maturing girls. *Dev Psychol.* 2011;47:1133–40.
  15. Deardorff J, Hayward C, Wilson KA, Bryson S, Hammer LD, Agras S. Puberty and gender interact to predict social anxiety symptoms in early adolescence. *J Adolesc Health.* 2007;41:102–4.
  16. Bakker MP, Ormel J, Lindenberg S, Verhulst FC, Oldehinkel AJ. Generation of interpersonal stressful events: the role of poor social skills and early physical maturation in young adolescents—the TRAILS study. *J Early Adolesc.* 2011;31:633–55.
  17. Haynie DL. Contexts of risk? Explaining the link between girls' pubertal development and their delinquency involvement. *Soc Forces.* 2003;82:355–97.
  18. Lynne SD, Graber JA, Nichols TR, Brooks-Gunn J, Botvin GJ. Links between pubertal timing, peer influences, and externalizing behaviors among urban students followed through middle school. *J Adolesc Health.* 2007;40:181.e7–181.e13.
  19. Mrug S, Elliott M, Gilliland MJ, Grunbaum JA, Tortolero SR, Cuccaro P, et al. Positive parenting and early puberty in girls: protective effects against aggressive behavior. *Arch Pediatr Adolesc Med.* 2008;162:781–6.
  20. Ullsperger JM, Nikolas MA. A meta-analytic review of the association between pubertal timing and psychopathology in adolescence: Are there sex differences in risk? *Psychol Bull.* 2017;143:903–38.
  21. Hamlat EJ, Snyder HR, Young JF, Hankin BL. Pubertal timing as a transdiagnostic risk for psychopathology in youth. *Clin Psychol Sci.* 2019;7:411–29.
  22. Hankin BL, Abramson LY, Moffitt TE, Silva PA, McGee R, Angell KE. Development of depression from preadolescence to young adulthood: emerging gender differences in a 10-year longitudinal study. *J Abnorm Psychol.* 1998;107:128–40.
  23. Graber JA. Pubertal timing and the development of psychopathology in adolescence and beyond. *Horm Behav.* 2013;64:262–9.
  24. Black SR, Klein DN. Early menarcheal age and risk for later depressive symptomatology: the role of childhood depressive symptoms. *J Youth Adolesc.* 2012;41:1142–50.
  25. Ge X, Conger RD, Elder Jr GH. Coming of age too early: Pubertal influences on girls' vulnerability to psychological distress. *Child Dev.* 1996;67:3386–400.
  26. Joinson C, Heron J, Lewis G, Croudace T, Araya R. Timing of menarche and depressive symptoms in adolescent girls from a UK cohort. *Br J Psychiatry.* 2011;198:17–23.
  27. Lam TH, Stewart SM, Leung GM, Lee PW, Wong JP, Ho LM, et al. Depressive symptoms among Hong Kong adolescents: Relation to atypical sexual feelings and behaviors, gender dissatisfaction, pubertal timing, and family and peer relationships. *Arch Sex Behav.* 2004;33:487–96.
  28. Kaltiala-Heino R, Kosunen E, Rimpelä M. Pubertal timing, sexual behaviour and self-reported depression in middle adolescence. *J Adolesc.* 2003;26:531–45.
  29. Kaltiala-Heino R, Marttunen M, Rantanen P, Rimpelä M. Early puberty is associated with mental health problems in middle adolescence. *Soc Sci Med.* 2003;57:1055–64.
  30. Rierdan J, Koff E. Depressive symptomatology among very early maturing girls. *J Youth Adolesc.* 1991;20:415–25.
  31. Stice E, Presnell K, Bearman SK. Relation of early menarche to depression, eating disorders, substance abuse, and comorbid psychopathology among adolescent girls.

- Dev Psychol. 2001;37:608–19.
32. Hayward C, Gotlib IH, Schraedley PK, Litt IF. Ethnic differences in the association between pubertal status and symptoms of depression in adolescent girls. *J Adolesc Health*. 1999;25:143–9.
  33. Carter R, Caldwell CH, Matusko N, Antonucci T, Jackson JS. Ethnicity, perceived pubertal timing, externalizing behaviors, and depressive symptoms among black adolescent girls. *J Youth Adolesc*. 2011;40:1394–406.
  34. Martino S, Lester D. Menarche and eating disorders. *Psychol Rep*. 2013;113(1):315–7.
  35. McGuire TC, McCormick KC, Koch MK, Mendle J. Pubertal maturation and trajectories of depression during early adolescence. *Front Psychol*. 2019;10:1362.
  36. Smith-Woolley E, Rimfeld K, Plomin R. Weak associations between pubertal development and psychiatric and behavioral problems. *Transl Psychiatry*. 2017;7:e1098.
  37. Toffol E, Koponen P, Luoto R, Partonen T. Pubertal timing, menstrual irregularity, and mental health: results of a population-based study. *Arch Womens Ment Health*. 2014;17:127–35.
  38. Joinson C, Heron J, Araya R, Lewis G. Early menarche and depressive symptoms from adolescence to young adulthood in a UK cohort. *J Am Acad Child Adolesc Psychiatry*. 2013;52:591–8.
  39. Angold A, Costello EJ, Erkanli A, Worthman CM. Pubertal changes in hormone levels and depression in girls. *Psychol Med*. 1999;29:1043–53.
  40. Balzer BW, Duke S-A, Hawke CI, Steinbeck KS. The effects of estradiol on mood and behavior in human female adolescents: a systematic review. *Eur J Pediatr*. 2015;174:289–98.
  41. Skovlund CW, Mørch LS, Kessing LV, Lidegaard Ø. Association of hormonal contraception with depression. *JAMA Psychiatry*. 2016;73:1154–62.
  42. Joinson C, Heron J, Araya R, Paus T, Croudace T, Rubin C, et al. Association between pubertal development and depressive symptoms in girls from a UK cohort. *Psychol Med*. 2012;42:2579–89.
  43. Horvath G, Knopik VS, Marceau K. Polygenic influences on pubertal timing and tempo and depressive symptoms in boys and girls. *J Res Adolesc*. 2020;30:78–94.
  44. Hooper L, Ness AR, Smith GD. Antioxidant strategy for cardiovascular disease. *The Lancet*. 2001;357:1705–6.
  45. Bell JA, Carslake D, Wade KH, Richmond RC, Langdon RJ, Vincent EE, et al. Influence of puberty timing on adiposity and cardiometabolic traits: a Mendelian randomisation study. *PLoS Med*. 2018;15:e1002641.
  46. Hartwig FP, Bowden J, de Mola CL, Tovo-Rodrigues L, Smith GD, Horta BL. Body mass index and psychiatric disorders: a Mendelian randomization study. *Sci Rep*. 2016;6:1–11.
  47. Wray NR, Ripke S, Mattheisen M, Trzaskowski M, Byrne EM, Abdellaoui A, et al. Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nat Genet*. 2018;50:668–81.
  48. Tyrrell J, Mulugeta A, Wood AR, Zhou A, Beaumont RN, Tuke MA, et al. Using genetics to understand the causal influence of higher BMI on depression. *Int J Epidemiol*. 2019;48:834–48.
  49. Lien L, Haavet OR, Dalgard F. Do mental health and behavioural problems of early menarche persist into late adolescence? A three year follow-up study among adolescent girls in Oslo, Norway. *Soc Sci Med*. 2010;71:529–33.
  50. Davey Smith G, Ebrahim S. ‘Mendelian randomization’: can genetic epidemiology contribute to understanding environmental determinants of disease? *Int J Epidemiol*. 2003;32:1–22.
  51. Day FR, Thompson DJ, Helgason H, Chasman DI, Finucane H, Sulem P, et al.



- Genomic analyses identify hundreds of variants associated with age at menarche and support a role for puberty timing in cancer risk. *Nat Genet.* 2017;49:834–41.
52. Smith GD, Holmes MV, Davies NM, Ebrahim S. Mendel's laws, Mendelian randomization and causal inference in observational data: substantive and nomenclatural issues. *Eur J Epidemiol.* 2020;35:99–111.
  53. Lundblad MW, Jacobsen BK. The reproducibility of self-reported age at menarche: The Tromsø Study. *BMC Womens Health.* 2017;17:1–7.
  54. Lawlor DA, Harbord RM, Sterne JA, Timpson N, Davey Smith G. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Stat Med.* 2008;27:1133–63.
  55. Sequeira M-E, Lewis SJ, Bonilla C, Smith GD, Joinson C. Association of timing of menarche with depressive symptoms and depression in adolescence: Mendelian randomisation study. *Br J Psychiatry.* 2017;210:39–46.
  56. Magnus P, Birke C, Vejrup K, Haugan A, Alsaker E, Daltveit AK, et al. Cohort Profile Update: The Norwegian Mother and Child Cohort Study (MoBa). *Int J Epidemiol.* 2016;45:382–8.
  57. Lawlor DA, Tilling K, Davey Smith G. Triangulation in aetiological epidemiology. *Int J Epidemiol.* 2016;45:1866–86.
  58. Munafò MR, Smith GD. Robust research needs many lines of evidence. *Nature.* 2018;553:399–402.
  59. Munafò MR, Higgins JP, Smith GD. Triangulating evidence through the inclusion of genetically informed designs. *Cold Spring Harb Perspect Med.* 2021;11:040659.
  60. Magnus MC, Guyatt AL, Lawn RB, Wyss AB, Trajanoska K, Küpers LK, et al. Identifying potential causal effects of age at menarche: a Mendelian randomization phenome-wide association study. *BMC Med.* 2020;18:1–17.
  61. Munafò MR, Nosek BA, Bishop DV, Button KS, Chambers CD, Du Sert NP, et al. A manifesto for reproducible science. *Nat Hum Behav.* 2017;1:1–9.
  62. Angold A, Costello EJ, Messer SC, Pickles A. Development of a short questionnaire for use in epidemiological studies of depression in children and adolescents. *Int J Methods Psychiatr Res.* 1995;5:237–49.
  63. Sharp C, Goodyer IM, Croudace TJ. The Short Mood and Feelings Questionnaire (SMFQ): a unidimensional item response theory and categorical data factor analysis of self-report ratings from a community sample of 7-through 11-year-old children. *J Abnorm Child Psychol.* 2006;34:379–91.
  64. Birmaher B, Khetarpal S, Brent D, Cully M, Balach L, Kaufman J, et al. The Screen for Child Anxiety Related Emotional Disorders (SCARED): scale construction and psychometric characteristics. *J Am Acad Child Adolesc Psychiatry.* 1997;36:545–53.
  65. Silva RR, Alpert M, Pouget E, Silva V, Trosper S, Reyes K, et al. A rating scale for disruptive behavior disorders, based on the DSM-IV item pool. *Psychiatr Q.* 2005;76:327–39.
  66. Turner NL, Joinson CJ, Peters TJ, Wiles NJ, Lewis GH. Validity of the Short Mood and Feelings Questionnaire in late adolescence. *Psychol Assess.* 2014;26:752–62.
  67. Paltiel L, Anita H, Skjerden T, Harbak K, Bækken S, Kristin SN, et al. The biobank of the Norwegian Mother and Child Cohort Study—present status. *Nor Epidemiol.* 2014;24:1–2.
  68. Turner S, Armstrong LL, Bradford Y, Carlson CS, Crawford DC, Crenshaw AT, et al. Quality control procedures for genome-wide association studies. *Curr Protoc Hum Genet.* 2011;68:1–19.
  69. Hemani G, Zheng J, Elsworth B, Wade KH, Haberland V, Baird D, et al. The MR-Base platform supports systematic causal inference across the human phenome. *elife.* 2018;7:e34408.
  70. Hemani G, Tilling K, Smith GD. Orienting the causal relationship between imprecisely

- measured traits using GWAS summary data. *PLOS Genet.* 2017;13:e1007081.
71. Schmitz D, Ek WE, Berggren E, Höglund J, Karlsson T, Johansson Å. Genome-Wide Association Study of Estradiol Levels, and the Causal Effect of Estradiol on Bone Mineral Density. *J Clin Endocrinol Metab.* 2021;106:e4471–e4486.
  72. Felix JF, Bradfield JP, Monnereau C, van der Valk RJP, Stergiakouli E, Chesi A, et al. Genome-wide association analysis identifies three new susceptibility loci for childhood body mass index. *Hum Mol Genet.* 2016;25:389–403.
  73. Richardson TG, Sanderson E, Elsworth B, Tilling K, Smith GD. Use of genetic variation to separate the effects of early and later life adiposity on disease risk: mendelian randomisation study. *bmj.* 2020;369:m1203.
  74. Levey DF, Stein MB, Wendt FR, Pathak GA, Zhou H, Aslan M, et al. Bi-ancestral depression GWAS in the Million Veteran Program and meta-analysis in >1.2 million individuals highlight new therapeutic directions. *Nat Neurosci.* 2021;24:954–63.
  75. Hernan MA, Robins J. Causal Inference: What if. Boca Raton: Chapman & Hill/CRC. 2020.
  76. Komsta L, Novomestky F. Moments, cumulants, skewness, kurtosis and related tests. R Package Version. 2015;14.
  77. Zeileis A, Köll S, Graham N. Various versatile variances: An object-oriented implementation of clustered covariances in R. *J Stat Softw.* 2020;95:1–36.
  78. Lee JJ, Wedow R, Okbay A, Kong E, Maghzian O, Zacher M, et al. Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nat Genet.* 2018;50:1112–21.
  79. Inoue A, Solon G. Two-Sample Instrumental Variables Estimators. *Rev Econ Stat.* 2010;92:557–61.
  80. Davey Smith G, Hemani G. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Hum Mol Genet.* 2014;23:89–98.
  81. Verbanck M, Chen C-Y, Neale B, Do R. Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. *Nat Genet.* 2018;50:693–8.
  82. Burgess S, Bowden J, Fall T, Ingelsson E, Thompson SG. Sensitivity Analyses for Robust Causal Inference from Mendelian Randomization Analyses with Multiple Genetic Variants. *Epidemiol Camb Mass.* 2017;28:30–42.
  83. Burgess S, Foley CN, Allara E, Staley JR, Howson JM. A robust and efficient method for Mendelian randomization with hundreds of genetic variants. *Nat Commun.* 2020;11:1–11.
  84. Burgess S, Thompson SG. Multivariable Mendelian Randomization: The Use of Pleiotropic Genetic Variants to Estimate Causal Effects. *Am J Epidemiol.* 2015 15;181:251–60.
  85. Sanderson E, Spiller W, Bowden J. Testing and correcting for weak and pleiotropic instruments in two-sample multivariable Mendelian randomization. *Stat Med.* 2021;40:5434–52.
  86. Grant AJ, Burgess S. Pleiotropy robust methods for multivariable Mendelian randomization. *Stat Med.* 2021;40:5813–30.
  87. Lakens D. Equivalence Tests: A Practical Primer for t Tests, Correlations, and Meta-Analyses. *Soc Psychol Personal Sci.* 2017;8:355–62.
  88. Sanderson E, Macdonald-Wallis C, Davey Smith G. Negative control exposure studies in the presence of measurement error: implications for attempted effect estimate calibration. *Int J Epidemiol.* 2018;47:587–96.
  89. Skrivankova VW, Richmond RC, Woolf BA, Yarmolinsky J, Davies NM, Swanson SA, et al. Strengthening the Reporting of Observational Studies in Epidemiology using Mendelian Randomization: the STROBE-MR Statement. *JAMA.* 2021;326:1614–21.
  90. Skrivankova VW, Richmond RC, Woolf BA, Davies NM, Swanson SA, VanderWeele

- TJ, et al. Strengthening the reporting of observational studies in epidemiology using mendelian randomisation (STROBE-MR): explanation and elaboration. *bmj*. 2021;375:n2233.
91. Lakens D, Scheel AM, Isager PM. Equivalence Testing for Psychological Research: A Tutorial. *Adv Methods Pract Psychol Sci*. 2018 1;1:259–69.
  92. Simonsohn U. Small Telescopes: Detectability and the Evaluation of Replication Results. *Psychol Sci*. 2015;26:559–69.
  93. Brion M-JA, Shakhbazov K, Visscher PM. Calculating statistical power in Mendelian randomization studies. *Int J Epidemiol*. 2013;42:1497–501.
  94. Elwert F, Winship C. Endogenous Selection Bias: The Problem of Conditioning on a Collider Variable. *Annu Rev Sociol*. 2014;40:31–53.

Additional file 1.

Design table with a summary of the research questions, hypotheses, power analyses, analysis plan, and conditions for interpretation of all potential outcomes.

Additional file 2.

Supplementary methods with information about the: a) multiple imputation, b) inverse probability weighting, c) genotyping, and d) psychometric properties of symptom scales.

Additional file 3.

Psychometric properties of 8-year symptom scales (ordinal Cronbach's alphas).

Additional file 4.

Overview of variables in the study, with details about items, reporters, and processing.

Additional file 5.

Overview of variables included in the ALSPAC study and similar variables available in MoBa, for replication purposes.

Additional file 6.

Summary of the arrays and batches used in the genotyping of the whole Norwegian Mother, Father, and Child Cohort.

Additional file 7.

Supplementary code containing scripts for data preparation, power analyses, and the analyses on which primary inferences will be made for each hypothesis, in each aim.

Additional file 8.

Results of power analyses based on simulated data for the null hypothesis significance tests and the equivalence tests for hypotheses 1-8 (Fig. S1-S8).