European Journal of Epidemiology
https://doi.org/10.1007/s10654-019-00550-1

METHODS

Bias from self selection and loss to follow-up in prospective cohort studies

Guido Biele[1] · Kristin Gustavson[1] · Nikolai Olavi Czajkowski[1] · Roy Miodini Nilsen[1] · Ted Reichborn-Kjennerud[1] · Per Minor Magnus[1] · Camilla Stoltenberg[1] · Heidi Aase[1]

Received: 20 August 2018 / Accepted: 7 August 2019

Abstract

Self-selection into prospective cohort studies and loss to follow-up can cause biased exposure-outcome association estimates. Previous investigations illustrated that such biases can be small in large prospective cohort studies. The structural approach to selection bias shows that general statements about bias are not possible for studies that investigate multiple exposures and outcomes, and that inverse probability of participation weighting (IPPW) but not adjustment for participation predictors generally reduces bias from self-selection and loss to follow-up. We propose to substantiate assumptions in structural models of selection bias through calculation of genetic correlations coefficients between participation predictors, outcome, and exposure, and to estimate a lower bound for bias due to self-selection and loss to follow-up by comparing effect estimates from IPP weighted and unweighted analyses. This study used data from the Norwegian Mother and Child Cohort Study and the Medical Birth Registry of Norway. Using the example of risk factors for ADHD, we find that genetic correlations between participation predictors, exposures, and outcome suggest the presence of bias. The comparison of exposure-outcome associations from regressions with and without IPPW revealed meaningful deviations. Assessment of selection bias for entire multi-exposure multi-outcome cohort studies is not possible. Instead, it has to be assessed and controlled on a case-by-case basis.

Keywords Bias · Self selection · Loss to follow up · Cohort study · Inverse probability weighting · Bayesian estimation · Directed acyclic graphs · ADHD

Introduction

The complex etiology of many disorders and ethical considerations often preclude experimental approaches to identifying their causes [1]. When controlled experimentation is not possible, cohort studies can provide valuable insights [2]. Prospective cohort studies are particularly valuable, because participants enroll before the outcome of interest has occurred. However, participation in cohort studies depends on socio-economic factors [3]. When the study sample is not a random sample from the population, selection bias is possible [4]. Hence, recent research investigated bias in exposure-outcome association estimates from large population-based prospective cohort studies empirically, by comparing associations in the study sample with those in the target population [5–8]. A related study assessed bias due to loss to follow-up by comparing association estimates from inclusion and follow-up participants [9]. This empirical approach to detecting selection bias can only evaluate bias when exposure and outcome data for the complete target population is available.

The structural approach to selection bias uses directed acyclic graphs (DAGs [10]) to explain the manifestation of bias. It requires information about participation predictors, for example age and education, and their relationship with exposure and outcome. Selection bias manifests if participation or participation predictors are colliders on an open path between exposure and outcome [11]. Hernán et al. [4] showed that even when there is no direct path between participation predictors and outcome, common unobserved causes of participation predictors and outcome can lead to selection bias. This manifests if, in addition, the exposure

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10654-019-00550-1) contains supplementary material, which is available to authorized users.

✉ Guido Biele
guido.biele@fhi.no

[1] Norwegian Institute of Public Health, Oslo, Norway

Published online: 26 August 2019

causes participation or if it causes or shares a common cause with predictors of participation (see Fig. 1a, b). Selection bias can also manifest due to effect modification, i.e. when population subgroups have varying participation rates and varying exposure-outcome associations (see Fig. 1c).

Figure 1 highlights that bias due to self-selection and loss to follow-up depends on the relationship between a specific exposure, outcome, participation predictors, and potential unobserved causes. Therefore, the presence or absence of bias cannot be determined for an entire cohort study that measures different exposures and outcomes. Instead, it has to be determined for each exposure-outcome-pair. Acquiring information about associations that determine selection bias is non-trivial, because *unobserved* common causes of participation predictors and outcome are central.

Common causes can be of environmental [12, 13] or genetic nature. Without presuming that common genetic causes carry more weight than environmental factors, we propose to use the more widely reported genetic correlation coefficients (r_G) from twin [14] and genome wide association studies [15] as an indicator for common unobserved causes. For instance, Bulik-Sullivan et al. [16] report single nucleotide polymorphism (SNP) based genetic correlations of $r_{GSNP} = 0.01$ and 0.731 between education in adulthood and birth-weight or childhood IQ, respectively. Hence, if one uses a study sample that over-represents well-educated mothers to examine associations of maternal depression with birth weight or childhood IQ, the latter association is more likely biased.

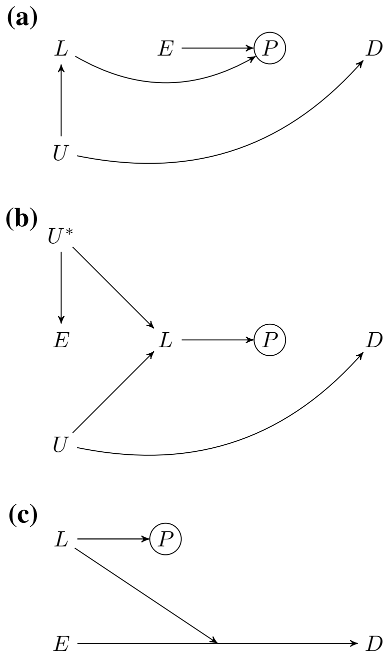


Fig. 1 Structural models of bias due to self-selection or loss to follow-up in prospective cohort studies. A spurious association manifests when participation P or participation predictors L are colliders on the path between exposure E and outcome D. P indicates conditioning on P, which opens a collider, resulting in selection bias. **a** Depicts a situation where L and E are independent as long as there is no conditioning on P. Inverse probability of participation weighting (IPPW), direct standardization (DS) and multilevel regression and post-stratification (MRP), or adjusting for L (AR) reduce this type of selection bias. **b** When E and L share an unobserved common cause, selection bias can only be reduced with IPPW. **c** Depicts bias due to effect modification, which can manifest in the absence of unobserved causes or conditioning on a collider. IPPW, DS, and MRP reduce this type of selection bias. **a** and **b** Modified from [4]

The structural approach does not provide an estimate of selection bias–magnitude. Still, the comparison of association estimates obtained with and without correction for self-selection can serve as a lower bound estimate of bias. Selection bias can be reduced by adjusting for participation predictors (adjusted regression, AR), by direct standardization with respect to participation predictors (DS [17]) or multilevel regression and poststratification (MRP [18]), and by weighting participants according to the inverse participation-probability (IPPW [19]). While IPPW reduces all types of selection bias displayed in Fig. 1, provided participation can be predicted well, DS/MRP reduce bias due to effect modification and structural bias when the exposure does not cause or share a common cause with participation predictors (as in Fig. 1a, c). AR only reduces bias in the absence of effect modification, or when the exposure does not cause or share a common cause with participation predictors (as in Fig. 1a). AR, DS/MRP cannot reduce selection bias when the exposure causes or shares a common cause with participation predictors (as in Fig. 1b) because conditioning on a collider introduces bias [11]. One disadvantage of IPP weighting is that it relies on a correct specification of the selection model and sufficient data. Therefore, AR should be favored if it is certain that participation predictors do not predict or have a common cause with the exposure in the target population (as in Fig. 1a).

In sum, this article proposes to evaluate selection bias in two steps. First, assumptions about common causes in structural models of exposure-outcome association and study participation should be substantiated, for example with estimates of genetic correlations. Second, a lower bound of selection bias magnitude can be estimated by comparing association estimates from IPPW and non-weighted analyses. In the remainder of the article we use associations between child and parental characteristics at birth and preschoolers' Attention Deficit Hyperactivity Disorder (ADHD) symptoms at age three in the Norwegian Mother and Child Cohort Study (MoBa) as an example to demonstrate assessment of bias due to self-selection and loss to follow-up in a large prospective cohort study. We estimate the joint effects of self-selection and loss to follow-up by assessing bias in the study sample for which outcome data is available, because both biases are present in longitudinal studies. *Note that in the remainder of the article "bias" refers exclusively to bias due to self-selection and loss to follow-up, and not to any other type of bias.*

Methods

Target population and study sample

Study sample

MoBa is a prospective population-based pregnancy cohort study conducted by the Norwegian Institute of Public Health [20, 21]. Participating mothers from all over Norway were recruited during routine ultrasound assessment in week 17 or 18 of their pregnancy in the period from 1999 to 2009. 41% of the invited women consented to participation. The cohort now includes 114,500 children, 95,200 mothers and 75,200 fathers. The current study is based on version 9 of the quality-assured data files released for research on October 2015. The reported analyses also use data from the Medical Birth Registry (MBRN), a national health registry containing information about all births in Norway [22].

The current analysis uses data from the main inclusion period from January 2001 to December 2009, in which 94,373 mothers returned the first MoBa questionnaire around the 20th pregnancy week. Of these 55,763 (59%) also returned the 6th MoBa questionnaire (at child age 3 years). Table 1 shows the bivariate distribution of maternal age and education in the MoBa sample and the target population, i.e. women in Norway who gave birth in the sampling period.

Participation predictors: socioeconomic data about the target population

We obtained aggregated data about maternal age, educational level, and number of children for all women in the target population from Statistics Norway.

Statistical analysis

R scripts for all analysis steps are available at https://github.com/gbiele/IPW/tree/master/AnalysisBIPW.

A detailed description of the statistical methods is in the supplementary information.

Investigating unobserved common causes with LD score regression

We calculated genetic correlations between predictors of participation, exposures, and outcome from publicly available summary results of genome wide association studies (GWAS) using linkage disequilibrium (LD) score regressions [15]. Table S2 lists maternal phenotypes for which we obtained GWAS summary statistics. Maternal genetic correlations also inform about common causes of maternal and child phenotypes because mothers and their children share 50% of their genes.

Table 1 Percent of mothers split by age and education in study sample (n = 54,557) and background population (n = 510,556), as well as coverage (% participation) of population subgroups in MoBa

Group	Education	< 20	20–24	25–29	30–34	35–39	40–49	All
MoBa	Elementary	0.2	0.4	0.4	0.4	0.2	0.1	1.7
	High-school	0.3	6.2	9.9	8.7	4.1	0.6	29.9
	Bachelor	0	1.8	16.7	18.0	6.2	0.8	43.7
	Master	0	0.0	6.0	12.8	5.3	0.6	24.7
	All	0.6	8.4	33.1	39.9	15.9	2.1	100.0
Population	Elementary	2	5.6	5.1	3.7	1.9	0.5	18.7
	High-school	0.3	6.8	11.9	10.3	4.6	0.8	34.7
	Bachelor	0	1.9	13.1	15.0	6.2	1.0	37.2
	Master	0	0.0	1.9	4.7	2.3	0.4	9.3
	All	2.3	14.4	32.0	33.7	15.0	2.5	100.0
Coverage	Elementary	1.3	0.7	1.0	1.3	1.4	1.3	1.0
	High-school	12.4	10.3	9.4	9.5	10.1	9.1	9.7
	Bachelor	–	10.6	14.4	13.5	11.4	9.3	13.2
	Master	–	–	35.1	30.4	25.8	19.5	29.8
	All	2.8	6.6	11.6	13.3	11.9	9.3	11.3

Data for MoBa participants from MoBa and MBRN, population data from statistics Norway

Outcome, exposures, and adjustment variables

We calculated an ADHD symptom score by summing the responses (Not, Somewhat, or Very often true, coded as 0, 1, 2) to 11 questions about ADHD symptoms that mothers' answered when the child was around 3 years old. Three separate analyses examined the magnitude of selection bias when estimating the association of preschoolers' ADHD symptoms with (a) birth-related exposures, (b) parental use of legal drugs, and (c) parental mental health including use of illegal drugs.

Table 2 describes the variables used in the analyses. MoBa assessed parental mental health with short forms of the symptom checklist (SCL5, [23]), the lifetime history of depression questionnaire (LTH, [24]), and the ADHD Self-Report Scale (ASRS, [25]). MoBa measured use of illegal drugs (cannabis, ecstasy, amphetamines, cocaine; less than 0.01% indicated having used heroin) before or in the pregnancy with Likert scales. As a dimensional measure of illegal drug-use, we used ability scores from an item response theory analysis [26].

All presented analyses used participants for which at least 50% of the analysis variables were available. We created 20 multiply imputed data sets through multiple imputation by chained equations as implemented in the R package mi [27].

Table 2 Description of variables and their use

Type	Variable	Description	Source	Used in
Outcome				
	Child ADHD	SS ADHD symptom ratings	Q6	IAU
Exposures, birth-related				
	small f. gest. age	1 = among 5% lightest in gest. wk.	MBRN	IAU
	preterm	1 = birth week < 37	MBRN	IAU
	m. drug use	Used any drug 1 = yes, no = 0	Q1	IAU
	m. drug use score	Ability score from IRT model	Q1	IAU
	m. LTH	SS Lifetime History of Depression	Q1	IAU
	m. SCL5	SS Symptom Check list	Q1	IAU
	p. drug use	Used any drug 1 = yes, no = 0	QF	IAU
Exposures, parental use of legal drugs				
	p. drug use score	Ability score from IRT model	QF	IAU
	m. smoking	Smoking in pregn.: 1 = yes, 0 = no	Q1	IAU
	m. num. cigarettes	Cigarettes per day in pregnancy	Q1	IAU
	m. alc. freq.	In pregnancy, 3 ordered categories	Q1	IAU
	m. eff. units. alc.	effective units of alcohol	Q1	IAU
	p. smoking	Smoking, 1 = yes, 0 = no	QF	IAU
	p. num. cigarettes	Cigarettes per day	QF	IAU
	p. alc. freq.	Drinking per week	QF	IAU
	p. typ. units. alc.	Typical number of units alcohol	QF	IAU
Adjustment variables				
	child sex	0 = boys, 1 = girls	MBRN	IA
	m. BMI deviat.	log (BMI/mode of BMI)	Q1	IA
	m. ADHD	SS ASRS	Q6	IA
	p. ADHD	SS ASRS	QF	IA
	p. age	8 ordered categories	QF	IA
	p. education	4 ordered categories	QF	IA
Participation predictors				
	m. education	4 ordered categories	Q1	I_SA
	m. age	6 ordered categories	Q1	I_SA
	Parity	Number of children born	MBRN	I_SA

m. = maternal, p. = paternal, Q = MoBa questionnaires, Q1 = at pregnancy week 17, QF = for fathers' (week 20), Q6 = at child age 3, MBRN = Medical Birth Registry of Norway, SS = sum score. All continuous and count variables except parity scaled to a mean of zero and a standard deviation of one. I = used in IPPW model, A = AR model, U = UR model, I_S in selection model for IPPWs calculation, not for adjustment in IPPW model

IPPWs and bias estimation

We calculated stabilized inverse probability weights using tabulated data about education, age and number of children for all birth giving mothers in Norway in the sampling period of MoBa. Because we rely on tabulated data, we used a binomial regression to estimate participation probabilities for population subgroups. We used a Bayesian hierarchical regression with random intercepts and slopes for the effect of age in subgroups defined by education and parity, in order to estimate effects of age also in small sub-groups reliably.

The estimation of a lower bound for bias involves estimating an IPPW and alternative regression models. The IPPW model estimates weighted and adjusted exposure outcome associations. The AR model, following typical practice for the analysis of cohort studies, adjusts for participation predictors instead of using IPPWs. The unadjusted regression (UR) does no adjustment and uses no IPPWs (c.f. Tables 2 and S1). To account for covariation of regression weights, we fit the three models (IPPW, AR, UR) simultaneously in a Bayesian framework, and define regression weights for exposures in the AR and UR models as weights from the IPPW model plus a difference terms. Because ADHD sum scores are constraint between 0 and 22, we used a beta binomial regression model and report associations as average marginal effects (AMEs).

Based on this analysis model, we calculated the lower bound bias estimate as the difference between AMEs from the IPPW and AR or UR models, respectively. We standardised the lower bound bias by dividing the difference with either the standard deviation or the mean of the posterior distribution of the IPPW estimate (c.f. [5, 28]). The latter approach appeals to the intuition that bias is problematic if the comparison standard is known with high precision/certainty, whereas the former appeals to the intuition that bias is problematic if it has a large deviation from the comparison standard.

To test for bias, we check how much of the posterior distribution of the bias estimate lies within a region of practical equivalence (ROPE, dashed vertical lines in Fig. 3), i.e., a bias magnitude that is for practical purposes equivalent to zero [29]. Here, we consider values of less than 0.5 standardised AME differences as practically equivalent with zero. To obtain a measure of risk for bias we calculate the log of the ratio of the posterior distribution inside and outside the ROPE, $\log(RR_b)$. For example, a $\log(RR_b)$ of -1.6 (3) means that the lower bound bias estimate is five (20) times as likely to lie outside (inside) the ROPE.

Regression analyses were performed with custom models implemented in the probabilistic programming language Stan [30] and fit via RStan [31].

Results

Statistics Norway recorded 510,561 women who became mothers in the period from 2001 to 2009. In the same period, 94,373 mothers returned the first MoBa questionnaire. Of these, 55,763 also returned the sixth questionnaire, which was sent out when children were 3 years old. 54,557 returned questionnaires with fewer than 50% missing data among the variables of interest. The study sample used for the reported analysis constitutes around 11% of the target population.

Socio-demographic composition of study sample and population

Mothers with elementary school education or less constitute around 18.7% of the target population and 1.7% of the MoBa sample (c.f. Table 1). 16.6% of mothers in the target population were younger than 25, compared to around 9.1% in the study sample. Accordingly, the participation rates vary between population subgroups: 29.7% of mothers in the target population with a master's degree are in the study sample, and around 1% of mothers with elementary school education. For parity, the difference between study sample and target population is less pronounced. The percentages of women in the target population (study sample) who had previously 0, 1, 2, or 3 or more children are 41.8 (50.9), 36.3 (32.5), 16.1 (13.9), 2.8 (5.8), respectively. Hence, the study sample over-represents mothers of firstborn children and under-represents those with more than two children.

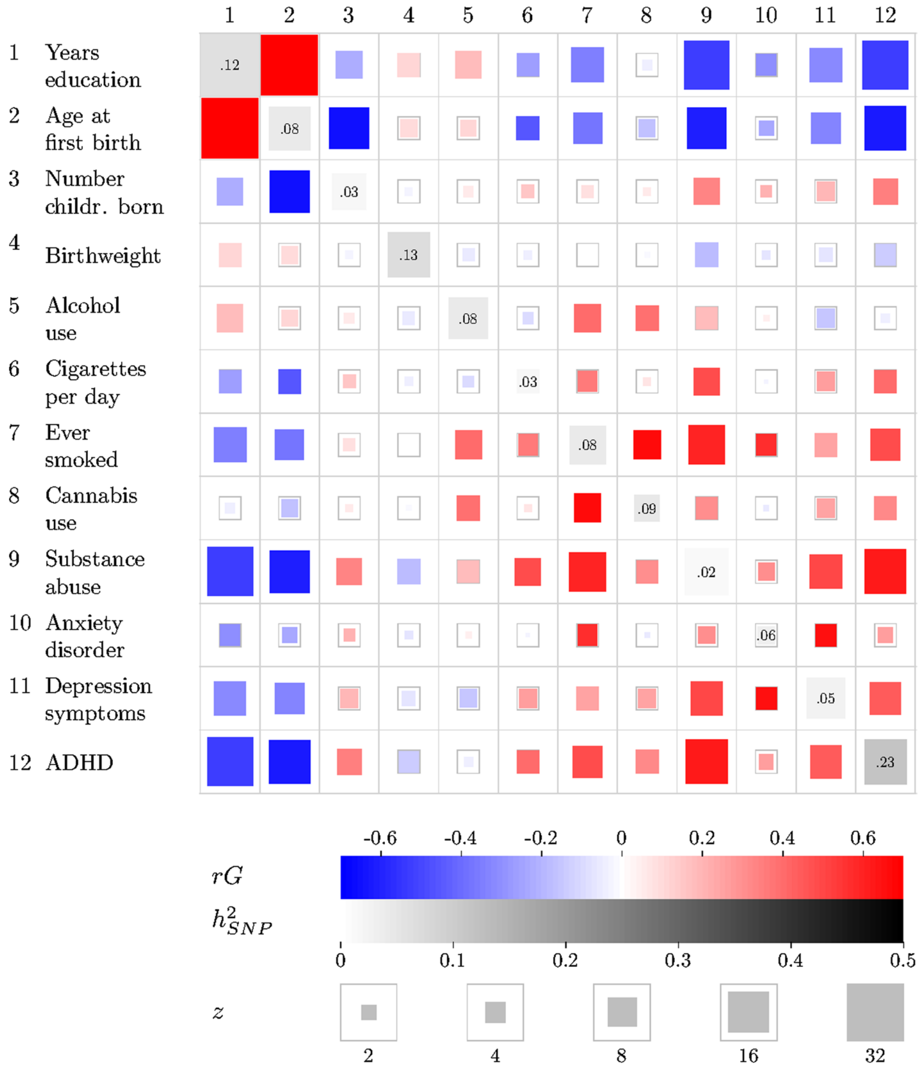
Inverse probability weights

The hierarchical binomial model captured participation well, as indicated by a correlation of $r = 0.99$ between modelled and observed participation rates (see Fig. S5). Mothers' education was the key variable to predict participation. Stabilized weights ranged between on average 0.31 and 21.36. The largest weights were for mothers with only elementary school education, and the smallest for mothers with a master's degree. We chose not to trim extreme weights, because this would result in improper weighting of the study sample. While some weights are very large, the associated population subgroups are typically represented with more than 100 children in the study sample.

Unobserved common causes

Genetic correlation results shown in Fig. 2 indicate unobserved common causes of participation predictors and outcome or exposures, respectively. For example, genes associated with "age at first birth" or "years of education" are also

Fig. 2 Genetic correlations as predictors of common unobserved causes. h^2_{SNP} = SNP based genetic heritability from LD score regression. r_G = genetic correlation between two traits based on LD score regression from publicly available GWAS summary statistics. Square colors indicate direction and size of correlations, the square size visualises z-values (which also depend on sample sizes). Gray square-outlines in the cells visualise the border to |z| = 4. The possibility of common causes of the participation predictor education the and outcome ADHD cannot be excluded. Education and some exposures like maternal depressive symptoms or smoking also appear to have common causes. Table S3 lists all genetic correlations and heritability estimates



(negatively) associated with ADHD, maternal mental health, or smoking. The estimated SNP based genetic heritability of the investigate phenotypes is typically not high, and often below 10%.

Selection bias for exposure-outcome associations

Figure 3 and Table 3 show AME estimates from an IPPW analysis (AME_{IPPW}), from an adjusted regression without weighting that adjusts for participation predictors (AME_{AR}), and from an unadjusted regression (AME_{UR}). Results from the IPPW analysis show, consistent with the literature, that most risk factors were positively, albeit weakly, associated with the ADHD symptom sum score. Maternal smoking had the strongest association: Mothers who indicated that they smoked, reported on average an ADHD symptom sum score for their child that was around 0.5 points higher (on a scale from 0 to 22) than mothers who indicated no smoking. Maternal drinking (frequency of alcohol use), maternal depressive symptoms (SCL5), and a low birth weight (small for gest. age) were also relatively strongly associated with ADHD sum scores. Associations with paternal variables other than drug use were generally weaker.

We estimated a lower bound of selection bias as the difference between average marginal effects (*AME*) from the IPPW and AR or UR models, standardised by either the mean or the standard deviation of the IPPW estimates. Figure 3 and Tables 3 and S4 show the results. For the AR model, only mean-standardised bias estimates for maternal depressive symptoms (SCL5, $\log(RR_b) = 4.5$) and maternal smoking in pregnancy (mSMOKE, $\log(RR_b) = 4.6$) are largely within the ROPE. Most bias estimates lie largely outside the ROPE. The risk ratio for having a bias larger than 0.5 for the AR model is higher than 20 (i.e. $\log(RR_b) < -3$) for 11 exposures when standardizing bias by the standard deviation of AME_{IPPW} and for 5 variables when standardizing by the mean of AME_{IPPW} (c.f. Table 3 and Figure S7).

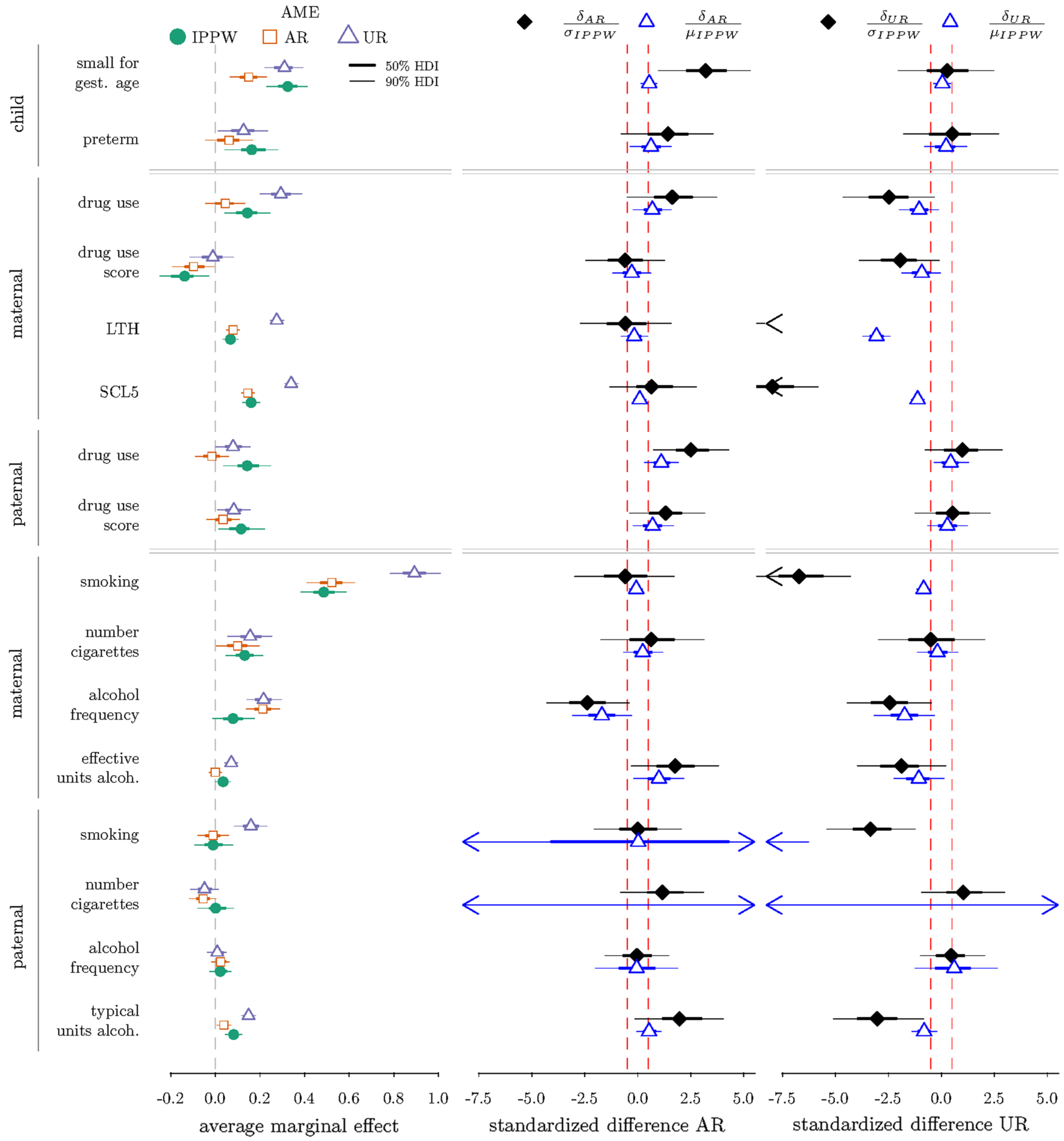


Fig. 3 Exposure-outcome associations and bias. Left panel: AME = Average Marginal Effect for a one unit increase of the exposure (all non-binary exposures z-standardised). Middle and right panel: δ_{AR} (δ_{UR}) are differences between estimates from adjusted (unadjusted) IPPW regressions, standardised by mean (μ_{IPPW}) or standard deviation (σ_{IPPW}) of the IPPW estimates. To confirm the absence of bias, the 90% HDI should fall between the dashed vertical lines, which enclose the region of practical equivalence with zero (ROPE). The ROPE contains standardised δs of less than 0.5. Bias estimates or HDIs outside the x-axis limits (− 7.5, + 5) are marked with < or > symbols

The results indicate both over- and under-estimation of associations in the AR analysis (e.g. frequency of maternal alcohol use and paternal drug use). IPPW and AR results also differ categorically, in that sometimes the IPPW results provide evidence for an association while the AR results do not (e.g. paternal drug use) and sometimes the opposite (paternal cigarettes per day).

Table 3 Means (m) and 90% highest density intervals (HDIs) of exposures outcome associations and standardised bias of AR results

	IPPW		AR		$\delta = \text{AR} - \text{IPPW}$		$\delta/\sigma_{\text{IPPW}}$			δ/μ_{IPPW}		
	m	HDI	m	HDI	m	HDI	m	HDI	log (RR)	m	HDI	log (RR)
Small f. gest. age	0.32	(0.23, 0.41)	0.15	(0.07, 0.23)	0.18	(0.05, 0.29)	3.21	(0.99, 5.30)	− 4.1	0.54	(0.17, 0.89)	− 0.3
Preterm	0.16	(0.04, 0.28)	0.06	(− 0.04, 0.17)	0.10	(− 0.05, 0.25)	1.42	(− 0.77, 3.55)	− 1.6	0.62	(− 0.34, 1.56)	− 0.5
m. drug use	0.14	(0.04, 0.24)	0.04	(− 0.04, 0.13)	0.10	(− 0.03, 0.23)	1.62	(− 0.48, 3.70)	− 1.8	0.69	(− 0.20, 1.57)	− 0.6
m. drug use score	− 0.14	(− 0.25, − 0.03)	− 0.10	(− 0.19, − 0.01)	− 0.04	(− 0.16, 0.08)	− 0.60	(− 2.43, 1.25)	− 0.9	− 0.29	(− 1.16, 0.59)	0.3
m. LTH	0.07	(0.04, 0.10)	0.08	(0.05, 0.11)	− 0.01	(− 0.05, 0.03)	− 0.59	(− 2.69, 1.56)	− 1.0	− 0.17	(− 0.78, 0.45)	1.2
m. SCL5	0.16	(0.12, 0.20)	0.15	(0.12, 0.17)	0.01	(− 0.03, 0.06)	0.65	(− 1.31, 2.74)	− 1.0	0.09	(− 0.18, 0.38)	4.5
p. drug use	0.14	(0.04, 0.25)	− 0.02	(− 0.09, 0.06)	0.16	(0.05, 0.27)	2.50	(0.76, 4.29)	− 3.5	1.11	(0.34, 1.90)	− 2.2
p. drug use score	0.12	(0.02, 0.22)	0.04	(− 0.04, 0.11)	0.08	(− 0.02, 0.19)	1.31	(− 0.39, 3.14)	− 1.5	0.70	(− 0.21, 1.66)	− 0.6
m. smoking	0.49	(0.39, 0.58)	0.52	(0.41, 0.62)	− 0.04	(− 0.18, 0.10)	− 0.60	(− 2.96, 1.71)	− 1.1	− 0.07	(− 0.37, 0.21)	4.6
m. num. cigarettes	0.13	(0.05, 0.21)	0.10	(0.01, 0.19)	0.03	(− 0.09, 0.15)	0.63	(− 1.75, 3.13)	− 1.2	0.23	(− 0.65, 1.17)	0.4
m. alc. freq.	0.08	(− 0.01, 0.17)	0.21	(0.14, 0.29)	− 0.13	(− 0.24, − 0.02)	− 2.38	(− 4.2, − 0.42)	− 3.0	− 1.69	(− 3.05, − 0.3)	− 2.5
m. eff. units alc.	0.03	(0.00, 0.07)	0.00	(− 0.03, 0.03)	0.03	(− 0.01, 0.07)	1.77	(− 0.29, 3.79)	− 2.0	1.00	(− 0.17, 2.14)	− 1.3
p. smoking	− 0.01	(− 0.09, 0.08)	− 0.01	(− 0.08, 0.06)	0.00	(− 0.10, 0.10)	0.00	(− 2.05, 2.03)	− 0.8	0.02	(− 10.3, 10.2)	− 2.7
p. num. cigarettes	0.00	(− 0.08, 0.08)	− 0.05	(− 0.12, 0.00)	0.06	(− 0.04, 0.15)	1.16	(− 0.80, 3.08)	− 1.4	41	(− 28.2, 109)	− 5.1
p. alc. freq.	0.02	(− 0.02, 0.07)	0.02	(− 0.01, 0.06)	− 0.00	(− 0.04, 0.04)	− 0.04	(− 1.54, 1.45)	− 0.3	− 0.05	(− 1.98, 1.87)	− 0.7
p. typ. units alc.	0.08	(0.05, 0.12)	0.04	(0.01, 0.07)	0.04	(− 0.00, 0.09)	1.97	(− 0.12, 4.02)	− 2.2	0.53	(− 0.03, 1.08)	− 0.2

IPPW, AR: Average marginal effects from IPPW, and AR models, respectively. σ_{IPPW} and μ_{IPPW} are standard deviation and mean of the posterior distribution of the IPPW regression coefficients. See Table S4 for statistics for the UR model

Discussion

Bias due to self-selection into studies and loss to follow-up is a threat to the validity of exposure-outcome association estimates from prospective cohort studies, because these often over-represent well educated, resourceful segments of the target population (Table 1, see also [3, 32, 33]). The structural approach to selection bias highlights that selection bias depends on the involved variables [4]. Therefore, it is not possible to evaluate selection bias generally for cohort studies that assess multiple exposures and outcomes. Among the statistical approaches to control bias, inverse probability of participation weighting (IPPW) is more generally able to correct bias than adjusted regression or direct standardization [4].

Using risk factor for ADHD as an example, we found that genetic correlations between participation predictors, exposures and outcome indicate potential bias, when maternal education predicts study participation. The analysis of associations between risk factors and ADHD in MoBa revealed substantial differences between association estimates obtained with IPPW and those obtained with adjustment for participation predictors (AR) or no adjustment (UR). There were only few instances of clear evidence against the presence of bias due to self-selection and loss to follow-up.

The current study reports more evidence for the presence of bias due to self-selection and loss to follow-up than previous investigations of large prospective cohort studies [5, 6, 9, 34]. Whereas previous reports used association estimates from the target population as a comparison standard for estimates from the study sample, this study used IPPW estimates. The validity of IPPW estimates as comparison standard depends on how well participation predictors predict participation [19]. In our study, the selection model predicted participation well. Another potential explanation for the stronger evidence for bias in our study is that bias in study samples at the onset of cohort studies is smaller because participation rates are higher. Further, because the heritability of ADHD is estimated to be higher compared to birth-related outcomes (c.f. Fig. 2 and [35]), selection bias due to common unobserved causes of participation predictors and outcome is expected to be larger for ADHD. Indeed, the strongest evidence for bias from earlier investigations comes from the association between maternal smoking and child ADHD [9]. Lastly, whereas previous studies evaluated bias by testing for a significant difference between sample and population estimates, equivalence testing [29, 36] is the proper approach to test if two association estimates are equal. Therefore, previous reports provided little statistical evidence for the absence of bias.

While the presented results indicate the presence of bias, one could reason that this is largely inconsequential, because the weighted and un-weighted association estimates typically point in the same direction. However, it is also important to recognise that in some cases the weighted and unweighted analyses led to categorically different conclusions. Crucially, in translational research, the magnitude of an association is important, so that not only non-detection of effects, but also errors in the estimation of effect sizes are problematic [37].

Conclusions about the presence or seriousness of bias can depend on how bias estimates are standardised or by how wide the ROPE is. Typically, bias estimates are standardised by the standard deviation of the unbiased parameter estimate [28], which we here replaced with the standard deviation of the corrected (IPPW) estimate. Similar to Nilsen et al. [5] we also estimated the percent deviation from the comparison standard, and found that this mean-standardised bias estimate was generally smaller. It is difficult to determine generally how large a bias is problematic. This should depend on the subject matter. We defined standardised deviations of less than 50% as practically equivalent with zero, which is considered to be a large effect [38], and still found clear evidence for bias.

Earlier assessments of bias in cohort studies that compared association estimates from study sample and target population are elegant in that their validity does not depend on assumptions about the causal relationship of exposure, outcome, and participation predictors. However, if population data about exposure and outcome are available, exposure-outcome associations need not be estimated from smaller study samples, and estimation of selection bias is superfluous. Using results from an IPPW regression as comparison standard rests on the assumption that the weighted study sample is a good representation of the target population, which is only the case if participation can be predicted well [19]. We found a high correspondence between predicted and observed participation rates, which suggests that the weighted MoBa sample represents the target population well. To verify that this test of the assumption is falsifiable, one can hypothesise a scenario that would have resulted in a violation. For example, if participation also strongly depended on maternal birth month, a selection model that uses only socio-demographic predictors would not predict participation well. Still, when calculating participation probabilities for population sub-groups, it remains possible that some bias results from within-group selection-bias, if there exist unmeasured participation predictors, that are independent of the measured predictors. This appears unlikely in the current analysis, because the key participation predictor education is strongly associated with unmeasured predictors like mental health.

The reliability of IPPW estimates also depends on the number of study-participants in population subgroups, especially if only few members of large population sub-groups participate in a study. However the low reliability of IPPW estimates in such circumstances, may indicate a weaknesses of the sampling strategy emploed for a study rather than a weakness of the IPPW. The IPPWs discussed in the current research are appropriate for relatively simple studies without time varying treatments or time to event analysis. For studies with such characteristics, more advanced weighting schemes like inverse probability of censoring weights (IPCW [39]) need to be employed.

While IPPW can remove bias due to self-selection and loss to follow-up, it cannot remove bias from unmeasured confounders. This is reflected in our finding of an association between maternal smoking and ADHD, which is likely due to familial genetic or environmental confounders [40]. It is important to emphasise that other biases than bias due to self-selection and loss to follow-up can still be present in estimates obtained with IPPW.

Structural analysis highlights that selection bias depends specifically on the involved variables, such that the presence or amount of bias for one association does not generalise to other associations. A first condition for selection bias to emerge is the presence of common unobserved causes of participation predictors, like education, and the outcome. A second condition is a direct or indirect causal relationship between participation predictors and the exposure. Figure 2 shows that genes (or associated environmental characteristics) can be common unobserved causes of mental health outcomes and the participation predictor education, and of education and exposures like smoking. It is therefore probable that non-weighted estimates of associations between e.g. maternal mental health, smoking or drinking behavior, and mental health related outcomes are biased in studies that over-represent certain educational groups. Still, the actual presence and magnitude of bias in such studies has to be examined on a case by case basis.

Structural analysis using directed acyclic graphs (DAGs) is a useful tool for the development of analysis strategies that remains underused. A practical argument against the use of DAGs is the uncertainty about hypothesised causal relationships. We proposed to use genetic correlation coefficients from LD score regression of publicly available GWAS summary statistics as one possibility to substantiate assumptions about unobserved common causes. The main motivation to focus on common genetic causes is the growing availability of GWAS summary statistics and methodological advances allowing estimation of heritability and genetic correlation coefficients from such statistics [15, 16]. Because GWAS studies are association studies, they do not provide unambiguous proof for a causal role of genes. Even if GWAS estimates are partly driven by environmental factors, genetic correlation estimates from GWAS summary statistics are of interest, because common unobserved environmental causes also contribute to the manifestation of selection bias. If direct estimates of common environmental causes are available, they should also be used for evaluating DAGs. Finally, the analysis of implied conditional independencies from competing causal models also allows examination of hypothesized causal relationships. As described in the supplementary discussion, such an analysis shows that the data analysed here are not consistent with a causal model in which exposure and participation predictors are independent in the target population (as in Fig. 1a).

A second challenge when using structural models is the difficulty of formulating DAGs for complex causal models [41]. When judging the presence of bias due to self-selection and loss to follow-up, a simple decision tree can supplant the formulation of a complete DAG, so that researchers can determine the potential for selection bias by answering a sequence of questions about the relationship of participation predictors, exposures, and outcomes. Figure 4 shows a decision tree that identifies when correction for bias is necessary, and what correction method can be used.

A topic closely related to selection bias is that of representativeness. While it was argued that representativeness can be detrimental to scientific inference, because understanding of mechanisms and careful control of relevant variables are central for this aim [42], others have emphasised the importance of representativeness—understood as the availability of weights for calculating valid population estimates [43]. Careful experimentation based on hypothesised mechanisms is undoubtedly central to scientific progress. Still, this approach does not describe the often-exploratory analyses of cohort study data well. Moreover, if one understands causal inference as the central goal of scientific inquiry, ignoring non-representativeness of unweighted study samples does not only undermine generalization to the population of interest, but can also lead to incorrect scientific inferences by facilitating the false discovery of associations, or prevent the detection of existing associations.

In conclusion, self-selection into cohort studies and loss to follow-up can lead to biased estimates of exposure-outcome associations from large population based cohort studies. Structural analysis and empirical results suggest that especially for mental health related exposures and outcomes selection bias is likely. Still, the dependency of bias on the specific outcome, exposure, and study participation predictors makes general statements about selection bias for multi-exposure multi-outcome studies impossible. Instead, each investigation of an exposure-outcome association has to assess selection bias. If bias is likely and valid participation predictors are available, weighting study participants by the inverse of their participation probability is a robust approach to control bias due to self-selection and loss to follow-up.

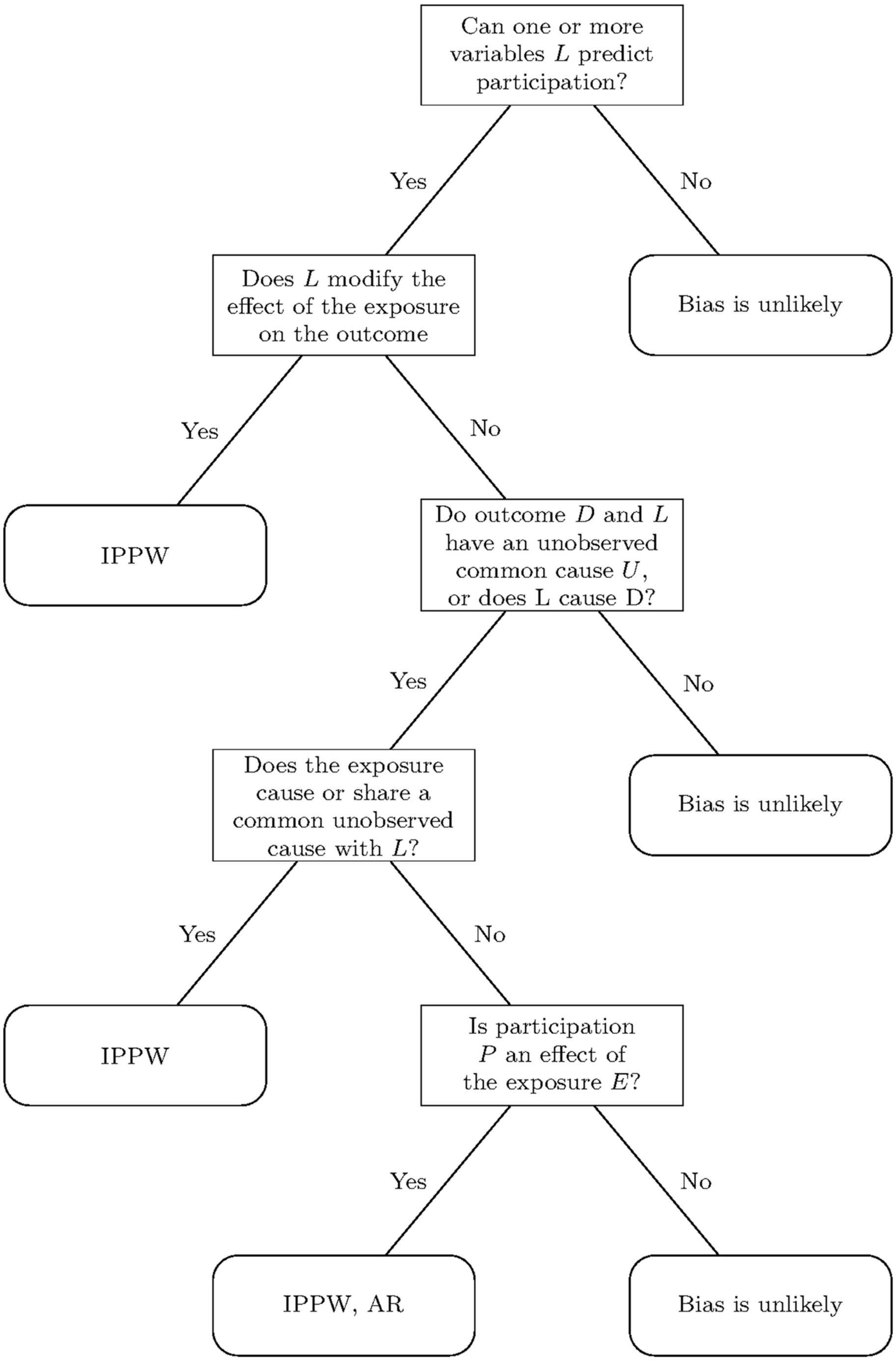

Can one or more variables L predict participation?
Yes
No
Does L modify the effect of the exposure on the outcome
Bias is unlikely
Yes
No
IPPW
Do outcome D and L have an unobserved common cause U, or does L cause D?
Yes
No
Does the exposure cause or share a common unobserved cause with L?
Bias is unlikely
Yes
No
IPPW
Is participation P an effect of the exposure E?
Yes
No
IPPW, AR
Bias is unlikely

Fig. 4 Decision tree for identification of selection bias and choice of approach to correct it. See Fig. 1 for causal diagrams that underlie the decision tree. To determine if selection bias is likely, and if so which correction method can be used, proceed through the questions from the top on. Ending in a node "Bias is unlikely" implies that an analysis without correction for selection bias likely results in estimates without selection bias. Otherwise, different correction methods can be used, depending on the underlying causal structure. IPPW stands for analysis with inverse probability of participation weighting, AR for adjusted regression. For reasons of brevity, this decision tree does not isolate cases where direct standardization (DS) or multilevel regression and post stratification (MRP) can correct bias

Acknowledgements The Norwegian Mother 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. The authors thank Eivind Ystrøm for discussing an earlier version of the research and the International Cannabis Consortium for providing GWAS summary statistics.

Funding This study was funded by the Norwegian Institute of Public Health.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Rothman KJ, Greenland S, Lash TL. Modern Epidemiology. Philadelphia: Lippincott Williams and Wilkins; 2008.
2. Greenland S. For and against methodologies: Some perspectives on recent causal and statistical inference debates. Eur J Epidemiol. 2017;32:3–20.
3. Galea S, Tracy M. Participation rates in epidemiologic studies. Ann Epidemiol. 2007;17:643–53.
4. Hernán MA, Hernández-Díaz S, Robins JM. A structural approach to selection bias. Epidemiology. 2004;15:615–25.
5. Nilsen RM, et al. Self-selection and bias in a large prospective pregnancy cohort in Norway. Paediatr Perinat Epidemiol. 2009;23:597–608.
6. Nohr EA, Frydenberg M, Henriksen TB, Olsen J. Does low participation in cohort studies induce bias? Epidemiology. 2006;17:413–8.
7. Nohr EA, Liew Z. How to investigate and adjust for selection bias in cohort studies. Acta Obstet Gynecol Scand. 2018;97:407–16.
8. Hatch EE, et al. Evaluation of selection bias in an internet-based study of pregnancy planners. Epidemiology. 2016;27:98–104.
9. Greene N, Greenland S, Olsen J, Nohr EA. Estimating bias from loss to followup in the Danish National Birth Cohort. Epidemiology. 2011;22:815–22.
10. Pearl J. Causal diagrams for empirical research. Biometrika. 1995;82:669–88.
11. Cole SR, et al. Illustrating bias due to conditioning on a collider. Int J Epidemiol. 2010;39:417–20.
12. Johnson W, et al. Does education confer a culture of healthy behavior? Smoking and drinking patterns in Danish twins. Am J Epidemiol. 2011;173:55–63.
13. Verweij KJH, Huizink AC, Agrawal A, Martin NG, Lynskey MT. Is the relationship between early-onset cannabis use and educational attainment causal or due to common liability? Drug Alcohol Depend. 2013;133:580–6.
14. Tambs K, et al. Genetic and environmental contributions to the relationship between education and anxiety disorders: A twin study. Acta Psychiatr Scand. 2012;125:203–12.
15. Bulik-Sullivan BK, et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat Genet. 2015;47:291–5.
16. Bulik-Sullivan BK, et al. An atlas of genetic correlations across human diseases and traits. Nat Genet. 2015;47:1236–41.
17. Miettinen OS. Standardization of risk ratios. Am J Epidemiol. 1972;96:383–8.
18. Downes M, et al. Multilevel regression and poststratification: A modelling approach to estimating population quantities from highly selected survey samples. Am J Epidemiol. 2018;187:1780–90.
19. Seaman SR, White IR. Review of inverse probability weighting for dealing with missing data. Stat Methods Med Res. 2013;22:278–95.
20. Magnus P, et al. Cohort profile: The Norwegian mother and child cohort study (MoBa). Int J Epidemiol. 2006;35:1146–50.
21. Magnus P, et al. Cohort profile update: The Norwegian mother and child cohort study (MoBa). Int J Epidemiol. 2016;45:382–8.
22. Irgens LM. The medical birth registry of Norway. Epidemiological research and surveillance throughout 30 years. Acta Obstet Gynecol Scand. 2000;79:435–9.
23. Tambs K, Moum T. How well can a few questionnaire items indicate anxiety and depression? Acta Psychiatr Scand. 1993;87:364–7.
24. Kendler KS, Neale MC, Kessler RC, Heath AC, Eaves LJ. The lifetime history of major depression in women. Reliability of diagnosis and heritability. Arch Gen Psychiatry. 1993;50:863–70.
25. Kessler RC, et al. Validity of the World Health Organization adult ADHD self-report scale (ASRS) screener in a representative sample of health plan members. Int J Methods Psychiatr Res. 2007;16:52–65.
26. Rizopoulos D. ltm: An R package for latent variable modelling and item response theory analyses. J Stat Softw. 2006;17:1–25.
27. Su Y-S, Gelman A, Hill J, Yajima M. Multiple imputation with diagnostics (mi) in R: Opening windows into the black box. J Stat Softw. 2011;45:1–31.
28. Austin PC. Some methods of propensity-score matching had superior performance to others: Results of an empirical investigation and Monte Carlo simulations. Biometr J. 2009;51:171–84.
29. Mascha EJ, Sessler DI. Equivalence and non-inferiority testing in regression models and repeated-measures designs. Anesth Analg. 2011;112:678–87.
30. Carpenter B, et al. Stan: A probabilistic programming language. J Stat Softw. 2017;76:1–29.
31. Stan Development Team. RStan: The R interface to Stan R package version 2.18.2. 2018. http://mc-stan.org/. Accessed 4 Feb 2019.
32. Vinther-Larsen M, et al. The Danish Youth Cohort: Characteristics of participants and non-participants and determinants of attrition. Scand J Public Health. 2010;38:648–56.
33. Howe LD, Tilling K, Galobardes B, Lawlor DA. Loss to follow-up in cohort studies: Bias in estimates of socioeconomic inequalities. Epidemiology. 2013;24:1–9.
34. Wolke D, et al. Selective drop-out in longitudinal studies and non-biased prediction of behaviour disorders. Br J Psychiatry. 2009;195:249–56.
35. Wu W, et al. The heritability of gestational age in a two-million member cohort: Implications for spontaneous preterm birth. Hum Genet. 2015;134:803–8.
36. Schuirmann DJ. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. J Pharmacokinet Biopharm. 1987;15:657–80.
37. Sullivan GM, Feinn R. Using effect size-or why the p value is not enough. J Grad Med Educ. 2012;4:279–82.
38. Cohen J. A power primer. Psycholog Bull. 1992;112:155–9.
39. Robins JM, Finkelstein DM. Correcting for noncompliance and dependent censoring in an AIDS Clinical Trial with inverse probability of censoring weighted (IPCW) log-rank tests. Biometrics. 2000;56:779–88.
40. Donovan SJ, Susser E. Commentary: advent of sibling designs. Int J Epidemiol. 2011;40:345–9.
41. Shrier I, Platt RW. Reducing bias through directed acyclic graphs. BMC Med Res Methodol. 2008;8:70.
42. Rothman KJ, Gallacher JEJ, Hatch EE. Why representativeness should be avoided. Int J Epidemiol. 2013;42:1012–4.
43. Keiding N, Louis TA. Perils and potentials of self-selected entry to epidemiological studies and surveys. J R Stat Soc Ser A (Stat Soc). 2016;179:319–76.