eAppendix

Loss to Follow-up in cohort studies: bias in estimates of socioeconomic inequalities

Laura D Howe, Kate Tilling, Bruna Galobardes, Debbie A Lawlor

Details of the Avon Longitudinal Study of Parents and Children (ALSPAC)

Pregnant women resident in one of three Bristol-based health districts with an expected date of delivery between 1-April 1991 and 31-December 1992 were eligible. Of the 14,541 women recruited (from an eligible population of approximately 20,000[1]; comparisons between the eligible and recruited population are presented in Supplementary Table 1), 13,988 children were alive at one year. Our analysis is limited to the 12,493 children, alive at one year, for whom data on maternal education are available. Ethical approval for the study was obtained from the ALSPAC Law and Ethics Committee and the Local Research Ethics Committees.

Measurement of outcomes for which data are available for (almost) the full cohort

1. Perinatal factors: birth length, birth weight, gestational age at delivery, breastfeeding:
Birth length (crown-heel) was measured by ALSPAC staff who visited newborns soon after birth (median 1 day, range 1-14 days), using a Harpenden Neonatometer (Holtain Ltd). Child's birth weight and gestational age were obtained from obstetric medical records. For all live births gestational age was as estimated by health care professionals in the medical records. Health care professionals used data from the woman's reported last menstrual period, paediatric assessment at birth, obstetric assessment during the antenatal period and ultrasound assessment; at the time that this cohort was established routine early pregnancy data scans were not conducted and it is likely that only a minority had gestational age determined by ultrasound scan. Breastfeeding was coded as 'any' or 'none' using information obtained from several questionnaires completed by mothers in the first six months after the infant's birth.

2. Maternal factors: maternal obesity, maternal smoking during pregnancy:
Maternal BMI was calculated using self-reported height and pre-pregnancy weight from a questionnaire administered at 12 weeks gestation. Maternal self-reported smoking in pregnancy were coded as 'any' or 'no' smoking at any time during pregnancy, using measures from several pregnancy questionnaires.

3. Child factors from routine data: Educational attainment at ages 11 and 14 years:
The UK education system is divided into a number of 'key stages'. Compulsory national tests mark the end of each key stage. Educational attainment data at key stages are available for all ALSPAC participants attending state schools (i.e. not those attending private schools) from linkage with the National Pupil Database. We utilise attainment data from exams taken at the end of key stage 2 (approximate age 11) and key stage 3 (approximate age 14). Scores in English, Mathematics, and Science tests were generated by summing all test scores in these subjects and converting to a percentage (with higher % scores indicating higher attainment). We used a summary measure of total educational attainment at both ages that was generated by combining the English, Mathematics and Science percentage scores and converting these

to a total percentage (again with 0% being the lowest possible attainment and 100% the highest possible).

Measurement of maternal education and other SEP indicators

Maternal education:
A questionnaire at 32 weeks gestation asked mothers to report their and their partner's educational attainment, which was categorised as below O-Level (Ordinary Level; exams taken in different subjects usually at age 15-16 at the completion of legally required school attendance, equivalent to today's UK General Certificate of Secondary Education), O-Level only, A-Level (Advanced-Level; exams taken in different subjects usually at age 18), or university degree or above.

Other SEP indicators:
The child's main caregiver was asked to report household income when the children were on average aged 3 and 4 years. These two measurements were averaged, housing benefit was added if appropriate for the household and incomes was equivalised to account for family size.[2] The Index of Multiple Deprivation (IMD) is a ward (administrative area)-based measure of poverty in the UK. It combines a range of domains (income, employment, health deprivation and disability, education skills and training, housing and geographical access to services) into a single deprivation score for each area; higher scores indicate more deprived areas.[3] Maternal age and parity were obtained from obstetric records. Household social class is measured as the highest of the mother's or her partner's occupational social class using data on job title and details of occupation collected about the mother and her partner from the mother's questionnaire at 32 weeks gestation. Social class is derived using the standard occupational classification (SOC) codes developed by the United Kingdom Office of Population Census and Surveys. Social class is categorised as I (professional), II (managerial and technical), III non-manual (skilled occupations, non-manual), III manual (skilled occupations, manual), IV (part skilled manual occupations) and V (unskilled manual occupations). In our analyses, armed forces were excluded since this represents a mixture of officers and lower rank staff. For Table 4 of the main manuscript, the social class variable was collapsed into a binary indicator of manual (classes III M, IV and V) or non-manual (classes I, II, III NM). Financial difficulties in affording food, clothes, heating and accommodation were reported by the mother in questionnaires at 32 weeks gestation. Mothers were asked, at the moment, how difficult they find it to afford each item, with possible answers of 'very difficult', 'fairly difficult', 'slightly difficult' or 'not difficult'. Access to a car by either the mother or her partner was reported in a questionnaire sent to the mother at 12 weeks gestation. Using information from the 12 week antenatal questionnaire, a crowding index was created by dividing the number of people in the household by the number of bedrooms. Whether or not the family is a single parent household was assessed from information the mother gave in the 12 week antenatal questionnaire about her partner and cohabitation status. Housing tenure was also assessed from the 12 week antenatal questionnaire, with mortgage and owned grouped into a home-owner category (72% of mothers) and all other responses grouped into a non-home-owner category.

Methods for multiple imputation:

We used switching regression in Stata as described by Royston.[3] We carried out 20 cycles of regression switching and generated 20 imputation datasets. The multiple imputation approach creates a number of copies of the data (in this case we generated 20 copies) in which missing values are imputed, with an appropriate level of randomness, by chained equations. [3] The main analysis results are obtained by averaging across the results from each of these 20 datasets using Rubin's rules and the procedure takes account of uncertainty in the imputation so that the standard errors for any regression coefficients (used to calculate p-values and 95% confidence intervals) take account of uncertainty in the imputations as well as uncertainty in the estimate. [3]

Methods for path analysis:

We used path analysis in Mplus to describe associations between maternal education, maternal smoking during pregnancy, birth weight, and participation at the 15 year clinic. Maternal education was included as a continuous variable, using the rank variable used to calculate SII and RIIs (see main text for details). Maternal smoking during pregnancy and participation at the 15 year clinic were binary variables. Birth weight was a continuous variable, standardised to have a mean of zero and variance of one. Each arrow in Figure 2 represents a linear regression – we mapped the binary indicator of maternal smoking during pregnancy to a standardised normal distribution, and as such the coefficients for this variable represent mean differences between non-smokers and smokers. Birth weight was regressed on maternal smoking in pregnancy and maternal education. Maternal smoking in pregnancy was regressed on maternal education. Attendance at the 15 year clinic was regressed on maternal education and maternal smoking during pregnancy. MLR estimation was used, which provides robust standard errors. Total effects for each regression are reported.

Methods for generating multidimensional SEP construct

Complete information on IMD, maternal and partner education, parity, maternal age, household social class, financial difficulties (food, clothes, heating and accommodation), car access, crowding, income, single parent status, and housing tenure for 8,210 women. Factor analysis was conducted on all of these indicators. The first factor explained 65% of the variance in the indicators. Each SEP indicator had a factor loading in the expected direction (worse area-deprivation score on IMD, higher parity, lower household occupational social class, higher crowding index, and single parent household associated with a reduction in SEP; all other indicators associated with an increase in SEP), and each SEP measure had a high 'uniqueness' score indicating that all were contributing to the multidimensional construct (**Supplementary table 4**). The index score was standardised (by subtracting the mean and dividing by the standard deviation) to generate a score with a mean of zero and variance of one.

eTable 1: Is non-participation related to outcomes?

Difference in educational attainment between the full eligible cohort identified in the National Pupil Database (i.e. those attending state schools but not private schools) and the children whose mothers consented to participate in the ALSPAC cohort.

	All eligible children identified in National Pupil Database,		**Mean difference (95% CI) comparing attendees with non-attendees (including all those identified in NPD as eligible to be ALSPAC participants)** **From linear regressions**
	n with data	*Mean (SD) in all children identified as eligible from NPD*	*Full ALSPAC cohort, N=12,493*
Summary score of educational attainment at age 11 (%)†	N=15,865	64.21 (16.05)	5.62 (5.10 to 6.14) P<0.001
Summary score of educational attainment at age 14 (%)†	N=13,613	50.90 (11.48)	3.83 (3.43 to 4.23) P<0.001

† Summary score of educational attainment is a sum of scores from tests in English, mathematics and science from compulsory school tests at ages 11 and 14 (Key Stages 2 and 3 of the British National Curriculum). The total score was converted to a percentage.

eTable 2: Socioeconomic patterning of loss to follow-up

	Full cohort	Attendees at age 10	% of original cohort retained	p value (comparison) with full cohort	Attendees at age 15	% of original cohort retained	p value (comparison) with full cohort
Maternal education, N(%)	N=12,493	N=7,045	56		N=5,075	41	
Less than O-Level	3,753 (30.0)	1,557 (22.4)	42		990 (19.5)	26	
O-Level	4,330 (34.7)	2,476 (35.2)	57		1,764 (34.8)	41	
A-Level	2,803 (22.4)	1,875 (26.6)	67		1,425 (28.1)	51	
Degree	1,607 (12.9)	1,117 (15.9)	70	<0.001	896 (17.7)	56	<0.001
Family income*	N=9,404	N=6,231	66		N=4,514	48	
Mean (SE)	5.64 (0.50)	5.69 (0.47)		<0.001	5.72 (0.46)		<0.001
Household occupational social class, N(%)	N=11,577	N=6,723	58		N=4,861	42	
IV and V	685 (5.9)	296 (4.4)	43		199 (4.1)	29	
IIIm	1,569 (13.6)	697 (10.4)	44		454 (9.3)	29	
IInm	2,947 (25.5)	1,663 (24.7)	56		1,155 (23.8)	39	
II	4,837 (41.8)	3,031 (45.1)	63		2,228 (45.8)	46	
I	1,539 (13.3)	1,036 (15.4)	67	<0.001	825 (17.0)	54	<0.001

*ln average family income at 33 and 47 months

Maternal education is defined as: O-Level (Ordinary Level; exams taken in different subjects usually at age 15-16 at the completion of legally required school attendance, equivalent to today's UK General Certificate of Secondary Education), O-Level only, A-Level (Advanced-Level; exams taken in different subjects usually at age 18), or university degree or above.

eTable 3: Is loss to follow-up related to outcomes?

	Full ALSPAC cohort (N=12,493)		Comparing attendees with non-attendees	
Outcomes on continuous scale			***Mean difference (95% CI)***	
	n with data	Mean (SD) or N (%)	Attendees at age 10, N=7,045	Attendees at age 15, N=5,075
Birth weight (g)	N=12,318	3406.68 (553.23)	23.75 (4.05 to 43.46) P=0.02	17.69 (-2.20 to 37.58) P=0.08
Birth length (cm)	N=9,655	50.63 (2.49)	0.07 (-0.03 to 0.17) P=0.16	0.10 (-0.01 to 0.20) P=0.06
Summary score of educational attainment at age 11 (%)†	N=10,365	66.16 (15.54)	7.26 (6.67 to 7.86) P<0.001	7.27 (6.68 to 7.86) P<0.001
Summary score of educational attainment at age 14 (%)†	N=8,856	52.24 (11.20)	4.61 (4.14 to 5.08) P<0.001	5.23 (4.77 to 5.68) P<0.001
Outcomes on binary scale			***Odds ratio (95% CI)***	
Maternal pre-pregnancy obesity (%)	N=10,466	571 (5.46%)	0.94 (0.79 to 1.11) P=0.46	0.81 (0.68 to 0.96) P=0.02
Maternal smoking during pregnancy (%)	N=12,152	2,947 (24.25%)	0.47 (0.43 to 0.51) P<0.001	0.45 (0.41 to 0.50) P<0.001
Preterm delivery (%)	N=12,492	728 (5.83%)	0.95 (0.81 to 1.10) P=0.46	0.88 (0.75 to 1.03) P=0.11
Never breastfed (%)	N=11,665	2,586 (22.17%)	0.44 (0.40 to 0.48) P<0.001	0.42 (0.38 to 0.47) P<0.001

† Summary score of educational attainment is a sum of scores from tests in English, mathematics and science from compulsory school tests at ages 11 and 14 (Key Stages 2 and 3 of the British National Curriculum). The total score was converted to a percentage. Educational attainment data are only available for those attending state schools; those attending private schools do not sit these national tests and so are not included in analyses

eTable 4: Risk of binary outcomes according to maternal education category amongst the full cohort, participants not lost to follow-up at age 10 and participants not lost to follow-up at age 15.

Maternal education	**% mothers obese**		
	Full cohort	Attendees at 10	Attendees at 15
less than O-level	8.3	8.6	8.7
O-level	5.4	5.5	4.9
A-level	4.0	4.0	3.8
degree or above	2.7	3.1	2.6
	% mothers who smoke		
	Full cohort	Attendees at 10	Attendees at 15
less than O-level	37.5	30.7	28.8
O-level	24.1	19.1	17.7
A-level	16.6	13.2	11.3
degree or above	8.1	7.8	7.3
	% preterm births		
	Full cohort	Attendees at 10	Attendees at 15
less than O-level	35.7	27.9	23.6
O-level	33.4	31.9	28.4
A-level	20.9	28.2	33.5
degree or above	10.0	12.0	14.6
	% never breastfed		
	Full cohort	Attendees at 10	Attendees at 15
less than O-level	50.5	42.3	37.9
O-level	36.2	39.9	42.5
A-level	11.4	15.5	16.3
degree or above	2.0	2.3	3.3

eTable 5: Assessing sensitivity of the results to the definition of maternal education:

Estimates of socioeconomic inequalities in outcomes with (almost) complete data amongst i) the full cohort, ii) participants who continue to participate at age 10 years, and iii) participants who continue to participate at age 15 years

Coefficients from linear or logistic regression comparing highest with lowest maternal education

Outcome	Full sample	Attendees at 10 years	Attendees at 15 years
Outcomes on a continuous scale, mean differences, null value = 0			
Birth weight (g)			
N	12,318	6,959	5,002
SII defining rank variable on full sample*	116.30 (79.78 to 152.82)	92.92 (44.55 to 141.30)	61.92 (4.83 to 119.00)
SII re-defining rank variable on each sample**	116.30 (79.78 to 152.82)	87.13 (40.82 to 133.45)	57.99 (3.56 to 112.41)
Using maternal education***	31.15 (21.43 to 40.88)	25.24 (12.37 to 38.11)	16.84 (1.67 to 32.01)
Birth length (cm)			
N	9,655	5,613	4,047
SII defining rank variable on full sample*	0.53 (0.34 to 0.71)	0.42 (0.17 to 0.66)	0.34 (0.05 to 0.63)
SII re-defining rank variable on each sample**	0.53 (0.34 to 0.71)	0.39 (0.16 to 0.62)	0.31 (0.04 to 0.59)
Using maternal education***	0.14 (0.09 to 0.19)	0.11 (0.05 to 0.18)	0.09 (0.02 to 0.17)
Summary score of educational attainment at age 11 (%)†			
N	10,365	6,828	4,529
SII defining rank variable on full sample*	23.73 (22.69 to 24.78)	20.29 (19.03 to 21.55)	19.40 (17.94 to 20.86)
SII re-defining rank variable on each sample**	23.73 (22.69 to 24.78)	19.31 (18.11 to 20.52)	18.41 (17.01 to 19.80)
Using maternal education***	6.35 (6.07 to 6.28)	5.43 (5.09 to 5.75)	5.19 (4.80 to 5.58)
Summary score of educational attainment at age 14 (%)†			
N	8,856	5,379	3,875
SII defining rank variable on full sample*	16.81 (15.97 to 17.66)	15.17 (14.12 to 16.22)	14.25 (13.03 to 15.46)
SII re-defining	16.81	14.37	13.50

rank variable on each sample**	(15.97 to 17.66)	(13.37 to 15.37)	(12.34 to 14.66)
Using maternal education***	4.51 (4.28 to 4.73)	4.06 (3.78 to 4.34)	3.82 (3.49 to 4.14)
Outcomes on a binary scale, odds ratios, null value = 1			
Maternal obesity			
N	10,466	6,332	4,544
RII defining rank variable on full sample*	0.23 (0.16 to 0.32)	0.24 (0.16 to 0.38)	0.21 (0.12 to 0.36)
RII re-defining rank variable on each sample**	0.23 (0.16 to 0.32)	0.26 (0.17 to 0.39)	0.23 (0.14 to 0.38)
Using maternal education***	0.67 (0.61 to 0.74)	0.69 (0.61 to 0.77)	0.66 (0.57 to 0.76)
Maternal smoking during pregnancy			
N	12,152	6,944	5,004
RII defining rank variable on full sample*	0.11 (0.09 to 0.13)	0.13 (0.10 to 0.16)	0.12 (0.09 to 0.17)
RII re-defining rank variable on each sample**	0.11 (0.09 to 0.13)	0.14 (0.11 to 0.18)	0.14 (0.10 to 0.19)
Using maternal education***	0.55 (0.52 to 0.58)	0.57 (0.54 to 0.61)	0.57 (0.54 to 0.61)
Preterm birth			
N	12,492	7,045	5,075
RII defining rank variable on full sample*	0.59 (0.45 to 0.79)	0.63 (0.43 to 0.92)	0.84 (0.53 to 1.33)
RII re-defining rank variable on each sample**	0.59 (0.45 to 0.79)	0.65 (0.45 to 0.93)	0.86 (0.55 to 1.33)
Using maternal education***	0.87 (0.80 to 0.94)	0.88 (0.79 to 0.96)	0.95 (0.84 to 1.07)
Never breastfed			
N	11,665	6,878	4,936
RII defining rank variable on full sample*	0.039 (0.031 to 0.047)	0.047 (0.035 to 0.062)	0.049 (0.034 to 0.069)
RII re-defining rank variable on each sample**	0.039 (0.031 to 0.047)	0.057 (0.044 to 0.075)	0.057 (0.041 to 0.080)
Using maternal	0.42 (0.39 to 0.44)	0.44 (0.41 to 0.47)	0.44 (0.40 to 0.49)

education***

*the slope index of inequality is the mean difference between the individuals of highest and lowest maternal education on a hypothetical underlying scale of continuous maternal education, based on the proportion of individuals in each maternal education category. The relative index of inequality is the odds ratio comparing the individuals of highest and lowest maternal education on a hypothetical underlying scale of continuous maternal education, based on the proportion of individuals in each maternal education category. For these analyses, the variable defining ranking of maternal education was based on the proportions of participants in each maternal education category within the full sample. This analysis applies the socioeconomic distribution of the full cohort at baseline to all analyses, including those restricted to participants not lost to follow-up at later stages of the cohort.

** the slope index of inequality is the mean difference between the individuals of highest and lowest maternal education on a hypothetical underlying scale of continuous maternal education, based on the proportion of individuals in each maternal education category. The relative index of inequality is the odds ratio comparing the individuals of highest and lowest maternal education on a hypothetical underlying scale of continuous maternal education, based on the proportion of individuals in each maternal education category. For these analyses, the variable defining ranking of maternal education was re-defined for participants attending at 10 years and again for those attending at 15 years, based on the proportions of these participants within each maternal education category. This analysis redefines the socioeconomic distribution at each analysis restricted to participants not lost to follow-up at later stages of the cohort. In these analyses, the proportion of mothers in the degree education category is much higher than in the full cohort at baseline – these participants will therefore be assigned a lower SEP (further away from the maximum value of 1 in the variable denoting the proportion of participants with a lower value of maternal education) than in the previous analysis based on the proportions of mothers in each educational category at baseline. Likewise, the proportion of mothers in the lowest educational category (< O-Level) is much lower in participants not lost to follow-up, so in these analyses these women will also be assigned a lower SEP (closer to the lowest possible value of 0 in the variable denoting the proportion of participants with a lower value of maternal education).

***maternal education is coded 1 for the lowest of four categories (less than O-level) up to 4 for the highest of four categories (degree and above). The coefficient therefore represents the mean difference (continuous outcomes) or odds ratio (binary outcomes) associated with each one category increase in maternal education. This analysis does not account for the differing proportions of participants in each maternal education category – therefore each increasing category of maternal education is assumed to confer the same benefits, regardless of how rare or common that educational level is.

eTable 6: factor loadings for multidimensional SEP construct

Variable	Factor score	(uniqueness)	Scoring coefficient used in prediction equation for multi-dimensional SEP construct
IMD	-0.3707	0.8626	-0.06491
Maternal age	0.3546	0.8743	0.07665
Parity	-0.2011	0.9595	-0.03077
Maternal education	0.4883	0.7616	0.11395
Partner education	0.5229	0.7266	0.12146
***Household social class**	-0.5334	0.7155	-0.11866
****Financial difficulties: food**	0.6596	0.5649	0.16210
****Financial difficulties: clothes**	0.6800	0.5376	0.15744
****Financial difficulties: heating**	0.6813	0.5359	0.20985
****Financial difficulties: accommodation**	0.5886	0.6535	0.11392
Car access	0.3154	0.9005	0.05377
Crowding	-0.4930	0.7569	-0.13415
Log income	0.6121	0.6253	0.13746
Single parent household	-0.1580	0.9750	-0.02660
Housing tenure	0.4516	0.7960	0.09186

* high score indicates low social class (1=professional, 2=managerial and technical, 3=non-manual (skilled occupations, non-manual), 4=skilled occupations, manual, 5=part skilled manual occupations and 6=unskilled manual occupations).
** A higher score on the financial difficulties measures indicates fewer financial problems.

eTable 7: Socioeconomic patterning of loss to follow-up using the multidimensional SEP construct.

	Full cohort, N=7,970	**Participants at age 9, N=5497**	**Participants at age 15, N=4039**
Latent SEP variable, mean (SD) of standardised score	0.0000000002 (1)	0.12 (0.95) p<0.001*	0.19 (0.94) p<0.001*

*P value for difference between full cohort and cohorts restricted by loss to follow-up

Supplementary material references

Reference List

(1) Boyd A, Golding J, MacLeod JA et al. Cohort profile: The 'Children of the 90s'; the index offspring of the Avon Longitudinal Study of Parents and Children (ALSPAC). *Int J Epidemiol.* 2012;In Press.

(2) Gregg, P., Propper, C., and Washbrook, E. Understanding the relationship between parental income and multiple child outcomes: a decomposition analysis. CASEpapers 129. 2007. Centre for Analysis of Social Exclusion, London School of Economics and Political Science, London, UK, CASEpapers.

Ref Type: Report

(3) Department of the Environment, Transport and the Regions. Indices of Deprivation 2000. 2000. Department of the Environment, Transport and the Regions. Regeneration Research Summary, No. 31.

Ref Type: Report