Making the missing at random assumption more plausible: Evidence from the 1958 British birth cohort

Declarations of interest: none

ABSTRACT

Objective

Non-response is unavoidable in longitudinal surveys. The consequences are lower statistical power and the potential for bias. We implemented a systematic data-driven approach to identify predictors of non-response in the National Child Development Study (NCDS; 1958 British birth cohort). Such variables can help make the missing at random assumption more plausible, which has implications for the handling of missing data.

Study Design and Setting

We identified predictors of non-response using data from the 11 sweeps (birth to age 55) of the NCDS (n = 17,415), employing parametric regressions and the LASSO for variable selection.

Results

Disadvantaged socio-economic background in childhood, worse mental health and lower cognitive ability in early life, and lack of civic and social participation in adulthood were consistently associated with non-response. Using this information, along with other data from NCDS, we were able to replicate the "population distribution" of educational attainment and marital status (derived from external data), and the original distributions of key early life characteristics.

Conclusion

The identified predictors of non-response have the potential to improve the plausibility of the missing at random assumption. They can be straightforwardly used as "auxiliary variables" in analyses with principled methods to reduce bias due to missing data.

KEYWORDS

Cohort studies; Longitudinal data; Missing data; Multiple imputation; National Child Development Study; Non-response.

RUNNING TITLE

Handling non-response in the 1958 British birth cohort

WHAT IS NEW?

- We capitalised on the rich data and long follow-up of the 1958 British birth cohort and implemented a systematic data-driven approach to identify predictors of non-response at each sweep of data collection.
- This approach allowed us to identify predictors of non-response not previously reported in the literature.
- Our findings have implications for users of NCDS and other longitudinal surveys as the identified predictors of non-response have the potential to maximise the plausibility of the missing at random assumption, informing analyses using principled approaches for missing data handling in order to restore sample representativeness.
- Our findings also have the potential to inform survey practice to reduce non-response levels in future waves of the 1958 British birth cohort and other longitudinal surveys.

INTRODUCTION

Non-response is unavoidable in longitudinal surveys. The consequences are smaller samples due to attrition, lower statistical power and decreased representativeness compared to the originally intended target population. With some exceptions where complete case analysis is valid (1-3), in the majority of analyses of longitudinal data bias will occur if the implications of selection due to incompleteness are not formally addressed (4, 5). There is a broad interdisciplinary consensus that missing data should be dealt with using principled approaches and it has recently been argued that "complete-case analysis should be used with the same caution we ascribe to unadjusted estimates, as its validity relies on strong, often unrealistic assumptions" (6).

Rubin described three missing data generating mechanisms: i) missing completely at random (MCAR)**;** ii) Missing at random (MAR)**;** iii) missing not at random (MNAR) (3, 7, 8). MCAR implies that the probability of non-response does not depend on any variable (measured or unmeasured), or that there are no systematic differences between the observed and missing data. MCAR is partially testable, since we can examine whether variables available in our data are associated with missingness. MAR implies that systematic differences between the missing values and the observed values can be explained by observed data, or that given the observed data, the reasons for missingness do not depend on unobserved variables. With some exceptions for specific missing data patterns (9, 10) the MAR assumption is untestable (11). The third mechanism - MNAR - implies that that the observed data are insufficient to explain variation in the probability of missingness. MNAR is also untestable and methods to deal with this type of missing data generating mechanism rely heavily on further – usually distributional - assumptions (12).

Contextualising the 1958 British National Child Development Study (NCDS) within Rubin's framework, we know that the missing data generating mechanism is not MCAR as previous work (13, 14) has shown that various variables are associated with non-response. In practice, as is expected to be the case in the vast majority of longitudinal surveys, in most analyses employing NCDS the missing data

generating mechanism is MAR or MNAR. Since both are largely untestable and considering that flexible solutions and software are available that return valid estimates assuming MAR, a pertinent question is how we can make MAR more plausible. Principled approaches that deal with missingness, such as multiple imputation (MI), full information maximum likelihood (FIML) and inverse probability weighting (IPW) assume MAR and thus are more likely to meaningfully reduce bias if careful steps have been taken to improve its plausibility (6-8, 15, 16). In the missing data methodology literature it is accepted that making MAR more plausible can be achieved by employing "auxiliary" – not in the substantive model of interest – variables, either in the imputation phase of MI, directly in FIML analysis, or in the derivation of non-response weights (17, 18). Effective auxiliary variables are thought to be variables associated both with non-response and the variable subject to missingness, as well as variables strongly associated with the variable subject to missingness only, since the expectation is that if so, they will also be associated with its missing values (5). There is disagreement as to whether variables associated only with non-response/missingness constitute effective auxiliary variables, with some authors arguing in favour of their inclusion (17, 19) and others against (5).

We capitalise on the rich data available in NCDS and present a systematic data-driven approach to identify predictors of non-response in all available sweeps. This has the potential to make the MAR assumption more plausible in analyses of NCDS data as it will allow researchers to identify the subset of predictors of non-response that are also associated with their missingness-affected substantive variables of interest and use these as auxiliary variables. The identified set of predictors of non-response therefore represents the maximal pool of such variables from which to draw on an analysis-specific basis. We also investigate whether by using the identified predictors of non-response along with other information available in NCDS we are able to restore sample representativeness despite selective attrition. By developing a principled approach to handling missing data handling in NCDS and providing empirical tests of the performance of our method, we hope to inform other work in the field.

MATERIALS AND METHODS

Data

The NCDS (20) is a well-characterised birth cohort study, with 10 major follow-ups since birth. The initial sample of 17,415 individuals – consisting of all babies born in Great Britain in a single week in 1958 – was supplemented with migrants at ages 7, 11 and 16. The most recent follow-up was at age 55, with high quality prospective data on social, physical, and psychological phenotypes available at every sweep. In 2002, when respondents were 44-45 years old, a biomedical survey was conducted in more than 9,000 respondents. We used the Office for National Statistics Annual Population Survey (APS) (21) to obtain estimates of the population distribution of key demographic characteristics for those born in 1958 and residing in Great Britain in 2008.

Exposures - predictors of non-response

NCDS datasets from the sweeps up to age 50 deposited in the UK Data Service include a total of 17,412 variables that could potentially be considered as predictors of non-response. We excluded "routed" variables (questions asked only of cohort members who gave a specific response to a previous question), used summary measures of scales rather the individual constituent items and excluded all binary variable with prevalence <1 % (further details in Methods S1). This resulted in 587 variables that met the criteria for inclusion in the analysis. They cover all domains captured by the NCDS (20), including indicators of socio-economic position, demographic characteristics, health, health behaviour, educational attainment, cognitive ability, personality traits, disability, relationships, social and political participation, biomarkers and others. In addition to these variables we calculated a summary variable that captures, for each sweep separately, whether or not cohort members participated in all previous sweeps.

Outcomes

We used binary variables indicating non-response for each sweep of NCDS from age 7 onwards. We defined non-response as participants who did not take part in the survey, either because of refusal, the survey team not been being able to establish contact, or because contact was not attempted, for example because of long-term refusal (Table S1). We did not consider as non-response participants that have died or emigrated since our aim was to identify predictors of non-response and not of mortality or emigration. We view missing data analysis as an attempt to restore sample representativeness with respect to a well-defined target population. The target population of NCDS, and any other longitudinal survey, is dynamic, as changes occur for example due to mortality. Considering that the NCDS mortality rate is representative of the population (Figure 1 and Table S2), the target population in each sweep of NCDS needs to be adjusted accordingly to reflect these changes. With the exception of modelling mortality as an outcome of interest, including participants that have died in any form of missing data analysis within NCDS would be the equivalent of generalising estimates to a non-existent (immortal) target population. The NCDS target population at each age is therefore all people born in 1958 who are alive and living in Great Britain at this age.

Analytic strategy

Our objective was to identify the important predictors of sweep-specific non-response. In order to achieve this we employed a three-stage analytic strategy using the identified 587 eligible variables as input. We opted for a multi-stage approach since the majority of the 587 potential predictors of non-response were not complete, and imputing all these simultaneously was not feasible. The first two stages therefore used complete case analyses, but in the third stage we used MI to impute missing values in the predictors of non-response. At each stage we modelled non-response with a log binomial

model with robust standard errors (modified Poisson regression (22)) that returns risk ratios to avoid bias due to non-collapsibility of the odds ratio (23) as non-response after age 23 becomes more common (>20%). Non-response at each sweep was analysed separately throughout the three-stage procedure.

The three-stage approach can be summarised as follows for non-response at sweep *t*:

- Stage 1: Complete case univariable modified Poisson regressions of non-response at sweep *t* on each potential predictor of non-response at sweep 0 up to sweep $t - 1$. Retain predictors with $p < 0.05$.

- Stage 2: Complete case multivariable modified Poisson regressions of non-response at sweep *t* on all retained predictors at sweep 0, then separately on all retained predictors at sweep 1, up to all retained predictors at sweep $t - 1$. Retain predictors with $p < 0.05$.

- Stage 3: MI using all retained variables plus non-response at sweep *t* in the imputation model. MI multivariable modified Poisson regressions for all retained predictors at sweep 0, up to sweep $t - 1$, adjusted for predictors at all previous (but not subsequent) sweeps. Retain predictors with $p < 0.001$.

Stage 3 allowed us to compare predictors of non-response from all stages of the life course and identify the set that has the potential to maximise the plausibility of the MAR assumption for a given NCDS sweep. Estimating a series of models in which predictors of non-response at a given sweep were adjusted for predictors at previous (but not subsequent) sweeps preserves the temporal sequence of the life course information available in NCDS while avoiding overadjustment from conditioning on variables on the causal pathway between a given predictor and non-response. When considering non-response at sweep *t*, the number of models estimated was thus *t* (one for each sweep between 0 and $t - 1$). So, for example, when considering non-response at sweep 6 (age 42), six models were estimated. The first of these models predicted non-response at age 42 from variables at sweep 0

(birth) that were retained after Stages 1 and 2; the final of these models predicted non-response at age 42 from variables at sweep 5 (age 33) that were retained after Stages 1 and 2, while also adjusting for variables at sweeps between 0 and 4 that had been retained after Stages 1 and 2.

In addition to protecting from overadjustment this approach ensures the richest appropriate adjustment, since from the results of Stage 2 we know that these are all the variables from the 587 included in the analysis potentially associated with non-response at a given sweep. We note that this approach introduces a causal structure based on the temporal sequencing of predictors of non-response as they appear in the various sweeps of NCDS, which is not typical in applications where prediction is primarily of interest. The rationale that underlies our decision is influenced by the fact that variables from early sweeps are relatively "complete" and are therefore more suitable candidates as auxiliary variables, considering that our ultimate goal is to inform applied missing data analyses in NCDS.

In stage 3 we employed MI with chained equations (24-26) and generated 50 datasets with imputed values using the previously identified (from Stage 2) sweep-specific predictors of non-response in the imputation phase. MI was carried out for each outcome (i.e. non-response at each sweep) separately as different predictors for non-response at each sweep had been identified from Stage 2.

We relied on *p*-values for variable selection within our regression-based approach. We did not consider the magnitude of association, as this is scale dependent, which is of particular concern for continuous predictors of non-response. For categorical predictors, the magnitude of the risk ratio for a given category would be dependent on the choice of baseline category and, in addition, for binary or categorical predictors, spuriously large (but imprecisely estimated) risk ratios could result from very low (or high) prevalence categories, leading to false positive variable selection.

The above three-stage procedure was repeated considering non-response at each sweep in turn. We defined "consistent" predictors of non-response to be variables identified at Stage 3 as predictors of

non-response at 50% or more of the sweeps in which they were eligible to be considered. For example, a variable from sweep 3 (age 16) could potentially be associated with non-response in seven subsequent sweeps. If such a predictor was associated with non-response in 4 or more subsequent sweeps it was selected as a consistent predictor of non-response.

As a robustness check for variable selection, we also employed the Least Absolute Shrinkage and Selection Operator (LASSO) (27) at stage 2. Group LASSO was used to appropriately consider categorical variables within the procedure (28). Considering that the majority of the 587 variables are not complete, we did not employ the LASSO or any other machine learning algorithm for variable selection at Stage 3. To the best of our knowledge, we are not aware of existing theory, let alone software, that allows the combination of MI with the LASSO or other machine learning approaches. We have therefore opted to use the LASSO as a form of sensitivity analysis at Stage 2 where missingness is less of an issue since variables are allowed to compete with others from the same sweep. However, a Stage 3 sensitivity analysis was also conducted using the variables selected using the LASSO at Stage 2, but using modified Poisson regression as in the primary analysis. The LASSO procedure was undertaken using logistic regression as modified Poisson models were not available, and the optimal set of variables was selected according to the minimum cross-validation error. As LASSO results were very similar to those from modified Poisson regressions, we present the latter (LASSO estimates for sweeps 1 and 2 are presented in the Web Appendix).

Restoring sample representativeness

In order to investigate whether the predictors of non-response identified at Stage 3, used in conjunction with other data from NCDS, have the potential to restore sample representativeness in NCDS despite selective attrition, we compared estimates from participants at age 50 with the known population distribution of educational attainment and marital status derived from the APS in 2008.

We also investigated whether the original distributions of paternal social class at birth and cognitive ability at age 7 could be replicated using data from only respondents at age 55 (i.e. disregarding data from non-respondents at age 55).

All analyses were conducted in Stata 14–16 and using gglasso in R.

RESULTS

Non-response in NCDS

In Table 1 we present descriptive statistics of participation in the NCDS from birth to 55 years. As expected, participation drops with time, with notable sample size reductions being at age 23, the first sweep where the cohort members were responsible for participating in the survey instead of their parents, as well as at age 44 for the NCDS biomedical sweep. Of the 17,415 cohort members who participated in the first sweep, 4,497 (25.8%) have participated in all 11 sweeps, 5,765 (33.1%) displayed monotone missingness and 7,153 (41.1%) exhibited non-monotone missingness. Of all 18,558 cohort members, 11,232 (60.5%) have taken part in 7 or more sweeps,.

Predictors of non-response

In the Web Appendix we present the results of the variable selection process we employed to identify predictors of non-response for all NCDS sweeps (Figures S1 – S10 and Tables S3-S13). In Tables 2, 3 and 4 we present risk ratios and 95% confidence intervals from 20 "consistent" predictors of non-response across sweeps of NCDS (further details of variable derivation in Methods S2). Females, cohort members that took part in all previous sweeps and those with fewer persons per room were more likely to participate in NCDS. Disadvantaged social class at birth was associated with non-response in most adult sweeps, but not – or even inversely associated – until age 23, indicating that parents from less advantaged socio-economic backgrounds were more likely to participate in the survey, but their offspring were more likely to drop out. Cognitive ability at ages 7 and 11 was consistently associated with survey participation, whereas conduct problems at age 16 were consistently associated with non-response. In adult sweeps, a systematic pattern emerged, with social participation, voting and marriage/cohabitation being associated with participation in NCDS. Other predictors associated with non-response included early life social problems and never having drank

alcohol by age 16. Using the LASSO rather than modified Poisson regression at Stage 2 resulted in the selection of a greater number of variables (Table S14). However, once the modified Poisson Stage 3 was conducted using the LASSO-selected Stage 2 variables, the resultant final selection of variables differed little from that in the primary analysis (Tables S15 and S16 vs. S4 and S5).

Restoring sample representativeness

We then evaluated the performance of our missing data strategy by comparing estimates after MI to those using only respondents at a given age. In Figure 2 we present the prevalence of those with degree or equivalent in the APS and NCDS. The prevalence of "degree or equivalent" at age 50 is 24.3% (95% confidence interval (CI) 23.4%-25.1%) based on the 9783 participants that took part in NCDS at age 50. This is higher than expected in the population based on APS data (18.6% (95% CI 17.3%-20.0%)or 18.9% (95% CI 17.4%-20.4%), depending on the inclusion of those born outside Great Britain), indicating that those with higher educational qualifications tend to drop out less from the survey on average. However, the estimate after MI from 15,806 NCDS participants alive and residing in Britain is 19.2% (95% CI 18.5%-19.9%), with a confidence interval which includes the estimates using APS data. Sample representativeness relative to APS estimates could similarly be restored for the prevalence of "no educational qualifications" (Figure S11) and for marital status (single and never married, Figure S12). Furthermore, we replicated the original distributions of paternal social class at birth (Figure S13) and cognitive ability at age 7 (Figure S14). In all cases examined, application of our missing data strategy therefore resulted in estimates that replicated external population benchmarks or full NCDS sample estimates.

CONCLUSIONS

Summary of findings

We applied a systematic data-driven approach to identify the important predictors of non-response in all available sweeps of NCDS. Identification of such variables has the potential to make the MAR assumption more plausible in analyses of NCDS data as it will allow researchers to identify the subset of predictors of non-response that are also associated with their missingness-affected substantive variables of interest and use these as auxiliary variables.

We observed prospective associations with non-response in all sweeps. In agreement with the literature on non-response in longitudinal surveys we found that those from a disadvantaged socio-economic background and men were less likely to respond to NCDS and are therefore less represented in later sweeps of the survey (14, 29). It has been argued that those with more advantaged socio-economic status are likely to appreciate the utility of research and hence have higher propensity to respond. In accordance with existing literature (30), we have shown that the intention to move was associated with non-response in subsequent sweeps, a finding consistent with the evidence on the association between residential mobility and attrition (31). Similarly to associations reported in the 1946 British birth cohort, we also found that early life cognitive ability was associated with survey participation (32), a finding perhaps expected due to the well-known association between early life cognitive ability and educational attainment (33). Consistent with a previous follow up of NCDS (14) we found that early life mental health in the form of conduct problems experienced at age 16 was associated with non-response in most sweeps of NCDS. Mental health problems in childhood and adolescence are known to be associated with low educational attainment, unemployment, unstable family formation, and criminal offending (34, 35), mechanisms that may explain the observed association with non-response. In accordance with the existing literature, we also found those single or divorced/separated/widowed have a lower propensity to respond than do those married (30). As

expected, taking part in previous sweeps of NCDS was strongly associated with participation in all sweeps.

Our data driven approach allowed us to identify predictors of non-response not previously reported, at least within the context of British birth cohorts. Strong associations were found between dimensions of social capital and non-response. Social and civic participation in the form of membership in group activities such as union membership, voting and having a strong social support network were associated with survey participation Considering that participating in surveys can be thought of as a form of social participation itself, these findings may reflect an overall propensity for participating in activities that are perceived as beneficial for the common good.

We have shown that by employing the identified predictors of non-response and other variables from NCDS we were able to replicate the known population distribution of educational attainment and marital status obtained from the APS, as well as the original distributions of paternal social class at birth and cognitive ability at age 7. These findings imply that improving the plausibility of MAR with observed data alongside principled methods for missing data handling has the strong potential to restore/maintain sample representativeness and reduce bias. However, this approach is not in any sense a formal test for MAR or MNAR, and there likely are variables in NCDS for which we would not be able to replicate the known population distribution.

Strengths and limitations

Strengths of this study include the availability of a population-based sample with 55 years of follow-up from birth and the systematic data driven approach that allowed us to capitalise on the rich information available in NCDS. Most studies investigating the association between survey participants' characteristics and non-response in longitudinal surveys have relied on theory-driven approaches, usually limiting their analysis to socio-economic and demographic characteristics.

Limitations of this study are the unavailability of interviewer information that could be used to inform our models and the fact that despite the strong multivariable adjustment, NCDS is an observational study and variables unavailable in NCDS (and hence not included in our analysis) and/or measurement error could have biased our results. We note that as our goal is to identify potential auxiliary variables, the fact that an unobserved confounder could explain the association between a particular variable and non-response does not have major implications for missing data handling, since its influence would – at least partly – be captured by the observed data. However, it has implications for the substantive interpretation of our findings on the influences of non-response, as different predictors may have been selected. Furthermore, our results can only be generalised to those born in 1958 in Britain or close to that year. In future work we plan to address these limitations by additionally considering information from administrative data linkages and polygenic risk scores, which have been shown to be associated with attrition (36), and to extend our analysis to younger cohorts such as the 1970 British Cohort Study, Next Steps and the Millennium Cohort Study to investigate generational differences in predictors of non-response. A further limitation of our overall approach to missing data handling, stemming from disagreement in the existing literature, regards the extent to which variables associated only with non-response/missingness and not with the missingness-affected substantive variables constitute effective auxiliary variables (5, 17, 19). We did not seek to address this question here – further research is required.

Implications for missing data analysis in NCDS

Our findings have implications for missing data handling in NCDS and have the potential to inform analyses in other longitudinal surveys. Although complete case analysis is known to return unbiased results in some scenarios, even when the data are not MCAR (1, 3), in the majority of analyses of NCDS, where missingness affects the exposure, outcome and potential confounders, a principled method would have to be employed to correct for missing data. The identified predictors of non-response

have the potential to be used as auxiliary variables in addition to the variables of substantive interest to the researcher in order to improve the plausibility of MAR in their analysis, especially if they are also associated with their variable(s) of interest that are subject to missingness. The inclusion of the identified predictors of non-response as auxiliary variables is straightforward in the imputation phase of MI and under somewhat more stringent distributional assumptions in FIML. They can also be used for the construction of weights that can be used in IPW analysis or analyses where MI and IPW are combined (18, 37, 38). However, while seeking to improve the plausibility of the MAR assumption in this manner is important, this does not mean that researchers should not consider whether their data are likely to be MNAR and conduct sensitivity analyses as appropriate (5, 39).

A publicly available step-by-step user guide based on our results is available on the CLS website to allow users of NCDS data to appropriately account for missing data (39). Associations between early life characteristics and non-response in adult sweeps are of similar strength to associations between adult characteristics and non-response. Since variables from the early sweeps of NCDS are generally affected much less by non-response, this implies that early life characteristics carry most of the information that improves the plausibility of MAR in NCDS.

Conclusion

Capitalising on the richness of NCDS we utilised a data-driven approach to empirically identify predictors of non-response that can improve the plausibility of the MAR assumption and which can inform analyses using principled approaches for missing data handling and restore sample representativeness. Identifying strong predictors of non-response at various stages of the life course has also the potential to inform survey practice to reduce non-response levels in future sweeps of NCDS and other longitudinal surveys.

FUNDING

This work was supported by the Economic and Social Research Council (ES/M001660/1). The funder played no role in study design, in the collection, analysis and interpretation of data, in the writing of the report, or in the decision to submit the article for publication.

REFERENCES

1. Bartlett JW, Carpenter JR, Tilling K, Vansteelandt S. Improving upon the efficiency of complete case analysis when covariates are MNAR. Biostatistics. 2014;15(4):719-30.
2. Daniel RM, Kenward MG, Cousens SN, De Stavola BL. Using causal diagrams to guide analysis in missing data problems. Statistical Methods in Medical Research. 2012;21(3):243-56.
3. Hughes RA, Heron J, Tilling K, Sterne JAC. Accounting for missing data in statistical analyses: multiple imputation is not always the answer. 2019.
4. Sterne JAC, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ. 2009;338.
5. Carpenter J, Kenward M. Multiple imputation and its application: John Wiley & Sons; 2012.
6. Perkins NJ, Cole SR, Harel O, Tchetgen Tchetgen EJ, Sun B, Mitchell EM, et al. Principled Approaches to Missing Data in Epidemiologic Studies. American journal of epidemiology. 2018;187(3):568-75.
7. Little RJA, Rubin DB. The analysis of social-science data with missing values. Sociological Methods & Research. 1989;18(2-3):292-326.
8. Little RJA, Rubin DB. Statistical Analysis with Missing Data Second Edition ed. Chichester: Willey; 2002.
9. Mohan K, Pearl J, Tian J, editors. Graphical models for inference with missing data. Advances in neural information processing systems; 2013.
10. Robins JM, Gill RD. Non-response models for the analysis of non-monotone ignorable missing data. Stat Med. 1997;16(1-3):39-56.
11. Molenberghs G, Beunckens C, Sotto C, Kenward MG. Every missingness not at random model has a missingness at random counterpart with equal fit. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2008;70(2):371-88.
12. Muthen B, Asparouhov T, Hunter AM, Leuchter AF. Growth modeling with nonignorable dropout: alternative analyses of the STAR*D antidepressant trial. Psychological methods. 2011;16(1):17-33.
13. Hawkes D, Plewis I. Modelling non-response in the National Child Development Study. Journal of the Royal Statistical Society: Series A (Statistics in Society). 2006;169(3):479-91.
14. Atherton K, Fuller E, Shepherd P, Strachan DP, Power C. Loss and representativeness in a biomedical survey at age 45 years: 1958 British birth cohort. Journal of Epidemiology and Community Health. 2008;62(3):216-23.
15. Enders CK. The performance of the full information maximum likelihood estimator in multiple regression models with missing data. Educational and Psychological Measurement. 2001;61(5):713-40.
16. Wooldridge JM. Inverse probability weighted estimation for general missing data problems. Journal of Econometrics. 2007;141(2):1281-301.
17. Enders CE. Applied missing data analysis. New York: Guilford; 2010.
18. Seaman SR, White IR. Review of inverse probability weighting for dealing with missing data. Stat Methods Med Res. 2011;22(3):278-95.
19. Collins LM, Schafer JL, Kam C-M. A comparison of inclusive and restrictive strategies in modern missing data procedures. Psychol Methods. 2001;6.
20. Power C, Elliott J. Cohort profile: 1958 British Birth Cohort (National Child Development Study). International Journal of Epidemiology. 2006;35(1):34-41.
21. Division OfNSSS. Annual Population Survey, 2004-2017. In: Division OfNSSS, editor. 13th Edition ed. London: UK Data Service; 2004 - 2017.
22. Zou G. A Modified Poisson Regression Approach to Prospective Studies with Binary Data. American Journal of Epidemiology. 2004;159(7):702-6.
23. Pang M, Kaufman JS, Platt RW. Studying noncollapsibility of the odds ratio with marginal structural and logistic regression models. Stat Methods Med Res. 2016;25(5):1925-37.

24. Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple Imputation by Chained Equations: What is it and how does it work? International journal of methods in psychiatric research. 2011;20(1):40-9.
25. Harel O, Mitchell EM, Perkins NJ, Cole SR, Tchetgen Tchetgen EJ, Sun B, et al. Multiple Imputation for Incomplete Data in Epidemiologic Studies. American journal of epidemiology. 2018;187(3):576-84.
26. White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. Stat Med. 2011;30.
27. Hastie T, Qian JJRJ. Glmnet vignette. 2014;9(2016):1-30.
28. Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2006;68(1):49-67.
29. Watson D. Sample attrition between waves 1 and 5 in the European Community Household Panel. European Sociological Review. 2003;19(4):361-78.
30. Watson N, Wooden M. Identifying factors affecting longitudinal survey response. Methodology of longitudinal surveys. 2009;1:157-82.
31. Plewis I, Ketende SC, Joshi H, Hughes G. The Contribution of Residential Mobility to Sample Loss in a Birth Cohort Study: Evidence from the First Two Waves of the UK Millennium Cohort Study. Journal of Official Statistics. 2008;24(3):365-85.
32. Stafford M, Black S, Shah I, Hardy R, Pierce M, Richards M, et al. Using a birth cohort to study ageing: representativeness and response rates in the National Survey of Health and Development. Eur J Ageing. 2013;10(2):145-57.
33. Sullivan A, Parsons S, Green F, Wiggins RD, Ploubidis G. The path from social origins to top jobs: social reproduction via education. The British journal of sociology. 2017.
34. Colman I, Murray J, Abbott RA, Maughan B, Kuh D, Croudace TJ, et al. Outcomes of conduct problems in adolescence: 40 year follow-up of national cohort. British Medical Journal. 2009;338.
35. Richards M, Abbott R. Childhood mental health and adult life chances in post-war Britain: insights from three national birth cohort studies. 2009.
36. Sallis H, Taylor AE, Munafò MR, Stergiakouli E, Euesden J, Davies NM, et al. Exploring the association of genetic factors with participation in the Avon Longitudinal Study of Parents and Children. International Journal of Epidemiology. 2018;47(4):1207-16.
37. Seaman SR, White IR, Copas AJ, Li L. Combining multiple imputation and inverse-probability weighting. Biometrics. 2012;68(1):129-37.
38. Sun B, Perkins NJ, Cole SR, Harel O, Mitchell EM, Schisterman EF, et al. Inverse-Probability-Weighted Estimation for Monotone and Nonmonotone Missing Data. American journal of epidemiology. 2018;187(3):585-91.
39. Silverwood R, Narayanan M, Dodgeon B, Ploubidis GB. Handling missing data in the National Child Development Study: User guide. London: UCL Centre for Longitudinal Studies. 2020.

TABLES

Table 1. Participation in the 1958 British National Child Development Study From Birth to 55 Years.

	Total cohort	Dead	Emigrants	Eligible sample	Participants	% of eligible sample
Birth - 1958	17638	0	0	17638	17415	98.7
Age 7 - 1965	18016[a]	821	475	16720	15425	92.3
Age 11 - 1969	18287[a]	840	701	16746	15337	91.6
Age 16 - 1974	18558[a]	873	799	16886	14654	86.8
Age 23 - 1981	18558	960	1196	16402	12537	76.4
Age 33 - 1991	18558	1049	1335	16174	11469	70.9
Age 42 - 2000	18558	1199	1268	16091	11419	71.0
Age 44 - 2002	18558	1321	1234	16003	9377	58.6
Age 46 - 2004	18558	1323	1272	15963	9534	59.7
Age 50 - 2008	18558	1459	1293	15806	9790	61.9
Age 55 - 2013	18558	1659	1286	15613	9137	58.5

[a] The original sample was supplemented by migrants born in 1958

Table 2. Estimated Risk Ratios and 95% Confidence Intervals for Consistent Predictors (Selected in at Least 50% of Possible Sweeps) of Non-response at Sweeps 1-5 (Ages 7-33) in the 1958 British National Child Development Study.

	Sweep 1 (age 7)		Sweep 2 (age 11)		Sweep 3 (age 16)		Sweep 4 (age 23)		Sweep 5 (age 33)	
	RR	95% CI	RR	95% CI	RR	95% CI	RR	95% CI	RR	95% CI
Non-response at previous sweep(s)										
Complete response	NA	NA	1.00	(reference)	1.00	(reference)	1.00	(reference)	1.00	(reference)
Incomplete response	NA	NA	5.76	5.28, 6.28	2.84	2.62, 3.06	2.10	1.99, 2.22	2.33	2.21, 2.46
Sweep 0 (age 0)										
Number of persons per room [per person]	1.10	1.05, 1.16	NS	NS	NS	NS	1.11	1.08, 1.14	1.11	1.09, 1.13
Sex of child										
Male	NS	NS	NS	NS	NS	NS	1.18	1.12, 1.25	1.22	1.16, 1.28
Female	NS	NS	NS	NS	NS	NS	1.00	(reference)	1.00	(reference)
Social class of mother's husband										
I	1.00	(reference)	NS	NS	NS	NS	1.00	(reference)	1.00	(reference)
II	0.66	0.51, 0.84	NS	NS	NS	NS	1.01	0.85, 1.21	1.06	0.90, 1.24
III non-manual	0.65	0.49, 0.86	NS	NS	NS	NS	0.91	0.75, 1.10	1.05	0.89, 1.25
III manual	0.59	0.47, 0.73	NS	NS	NS	NS	1.13	0.96, 1.32	1.21	1.04, 1.40
IV	0.72	0.57, 0.92	NS	NS	NS	NS	1.14	0.96, 1.36	1.30	1.11, 1.52
V	0.80	0.62, 1.02	NS	NS	NS	NS	1.46	1.23, 1.73	1.72	1.47, 2.00
Sweep 1 (age 7)										
Cognitive ability summary [per unit]	NA	NA	0.85	0.80, 0.91	NS	NS	0.86	0.83, 0.89	0.87	0.84, 0.89
Social problems (alcoholism etc.) [per problem]	NA	NA	NS	NS	NS	NS	NS	NS	1.10	1.07, 1.13
Sweep 2 (age 11)										
Cognitive ability summary [per 10 units]	NA	NA	NA	NA	NS	NS	0.91	0.88, 0.94	0.89	0.87, 0.92
Sweep 3 (age 16)										
Conduct problems [per unit]	NA	NA	NA	NA	NA	NA	1.10	1.07, 1.13	NS	NS
How long since child drank alcohol										
Less than 1 week	NA	NA	NA	NA	NA	NA	NS	NS	1.00	(reference)
2 to 4 weeks	NA	NA	NA	NA	NA	NA	NS	NS	0.97	0.88, 1.07

5+ weeks	NA	NA	NA	NA	NA	NA	NS	NS	1.04	0.95, 1.13
Do not remember	NA	NA	NA	NA	NA	NA	NS	NS	1.11	1.01, 1.22
Never had one	NA	NA	NA	NA	NA	NA	NS	NS	1.27	1.14, 1.41
Test 2 – mathematics comprehension [per 10 units]	NA	NA	NA	NA	NA	NA	NS	NS	0.82	0.76, 0.88
Sweep 4 (age 23)										
Voted in 1979 general election										
Didn't vote	NA	NA	NA	NA	NA	NA	NA	NA	1.24	1.17, 1.32
Voted	NA	NA	NA	NA	NA	NA	NA	NA	1.00	(reference)
Legal marital status										
Single	NA	NA	NA	NA	NA	NA	NA	NA	NS	NS
Married	NA	NA	NA	NA	NA	NA	NA	NA	NS	NS
Separated/divorced/widowed	NA	NA	NA	NA	NA	NA	NA	NA	NS	NS

NA: Not applicable (predictor variable observed concurrently with or subsequent to non-response variable); NS: Not selected.

Table 3. Estimated Risk Ratios and 95% Confidence Intervals for Consistent Predictors at Sweeps 0-4 (ages 0-23) (Selected in at Least 50% of Possible Sweeps) of Non-response at Sweeps 6-9 (ages 42-55) in the 1958 British National Child Development Study.

	Sweep 6 (age 42)		Biomedical sweep (age 44)		Sweep 7 (age 46)		Sweep 8 (age 50)		Sweep 9 (age 55)	
	RR	95% CI	RR	95% CI	RR	95% CI	RR	95% CI	RR	95% CI
Non-response at previous sweeps										
Complete response	1.00	(reference)	1.00	(reference)	1.00	(reference)	1.00	(reference)	1.00	(reference)
Incomplete response	3.83	3.57, 4.11	3.37	3.17, 3.58	7.17	6.53, 7.88	6.28	5.71, 6.91	5.93	5.39, 6.54
Sweep 0 (age 0)										
Number of persons per room [per person]	1.11	1.09, 1.13	1.08	1.07, 1.10	1.08	1.06, 1.10	1.07	1.05, 1.09	1.06	1.04, 1.08
Sex of child										
Male	1.19	1.13, 1.25	1.07	1.03, 1.11	1.14	1.10, 1.19	1.11	1.07, 1.46	1.13	1.09, 1.18
Female	1.00	(reference)	1.00	(reference)	1.00	(reference)	1.00	(reference)	1.00	(reference)
Social class of mother's husband										
I	1.00	(reference)	1.00	(reference)	1.00	(reference)	1.00	(reference)	1.00	(reference)
II	0.94	0.80, 1.11	1.08	0.95, 1.23	1.09	0.96, 1.25	0.98	0.86, 1.12	1.00	(reference)
III non-manual	1.02	0.86, 1.20	1.14	1.00, 1.30	1.13	0.99, 1.29	1.05	0.92, 1.20	1.11	1.01, 1.22
III manual	1.18	1.02, 1.36	1.25	1.12, 1.40	1.27	1.13, 1.43	1.18	1.05, 1.32	1.35	1.26, 1.43
IV	1.22	1.05, 1.43	1.32	1.16, 1.49	1.34	1.18, 1.52	1.27	1.12, 1.43	1.41	1.31, 1.53
V	1.51	1.30, 1.77	1.55	1.38, 1.75	1.62	1.43, 1.83	1.45	1.28, 1.63	1.69	1.57, 1.82
Sweep 1 (age 7)										
Cognitive ability summary [per unit]	0.83	0.80, 0.85	0.85	0.83, 0.87	0.83	0.81, 0.85	0.84	0.82, 0.86	0.82	0.80, 0.84
Social problems (alcoholism etc.) [per problem]	NS	NS	1.04	1.02, 1.06	1.03	1.01, 1.05	1.07	1.04, 1.09	1.04	1.02, 1.06
Sweep 2 (age 11)										
Cognitive ability summary [per 10 units]	0.88	0.85, 0.90	0.90	0.88, 0.92	0.89	0.88, 0.91	0.90	0.88, 0.92	0.88	0.86, 0.89
Sweep 3 (age 16)										
Conduct problems [per unit]	1.08	1.05, 1.11	1.06	1.04, 1.08	NS	NS	1.06	1.04, 1.08	1.05	1.03, 1.07
How long since child drank alcohol										
Less than 1 week	1.00	(reference)	1.00	(reference)	1.00	(reference)	1.00	(reference)	1.00	(reference)

2 to 4 weeks	1.05	0.96, 1.14	1.05	0.99, 1.12	1.06	0.99, 1.13	1.04	0.97, 1.11	1.03	0.96, 1.10
5+ weeks	1.08	1.00, 1.18	1.06	1.00, 1.14	1.09	1.02, 1.17	1.02	0.95, 1.10	1.04	0.97, 1.11
Do not remember	1.11	1.01, 1.23	1.14	1.06, 1.22	1.14	1.06, 1.23	1.12	1.04, 1.20	1.12	1.04, 1.19
Never had one	1.27	1.13, 1.42	1.21	1.11, 1.31	1.26	1.17, 1.37	1.21	1.10, 1.32	1.22	1.13, 1.31
Test 2 – mathematics comprehension [per 10 units]	NS	NS	0.90	0.85, 0.94	0.87	0.82, 0.92	0.88	0.83, 0.93	0.86	0.82, 0.90
Sweep 4 (age 23)										
Voted in 1979 general election										
Didn't vote	1.25	1.18, 1.33	1.13	1.08, 1.19	1.16	1.11, 1.22	1.18	1.13, 1.24	1.16	1.11, 1.21
Voted	1.00	(reference)	1.00	(reference)	1.00	(reference)	1.00	(reference)	1.00	(reference)
Legal marital status										
Single	1.05	0.97, 1.13	NS	NS	NS	NS	1.04	0.99, 1.10	1.12	1.03, 1.21
Married	1.00	(reference)	NS	NS	NS	NS	1.00	(reference)	1.00	(reference)
Separated/divorced/widowed	1.32	1.16, 1.51	NS	NS	NS	NS	1.21	1.09, 1.34	1.24	1.11, 1.38

NA: Not applicable (predictor variable observed concurrently with or subsequent to non-response variable); NS: Not selected.

Table 4. Estimated Risk Ratios and 95% Confidence Intervals for Consistent Predictors at Sweeps 5-8 (ages 33-50) (Selected in at Least 50% of Possible Sweeps) of Non-response at Sweeps 6-9 (ages 42-55) in the 1958 British National Child Development Study.

	Sweep 6 (age 42)		Biomedical sweep (age 44)		Sweep 7 (age 46)		Sweep 8 (age 50)		Sweep 9 (age 55)	
	RR	95% CI	RR	95% CI	RR	95% CI	RR	95% CI	RR	95% CI
Sweep 5 (age 33)										
Voted in 1987 general election										
Didn't vote	NS	NS	NS	NS	1.12	1.06, 1.19	1.16	1.10, 1.23	1.16	1.11, 1.21
Voted	NS	NS	NS	NS	1.00	(reference)	1.00	(reference)	1.00	(reference)
Social capital score (people turn to for advice, support) [per 10 units]	0.81	0.77, 0.85	0.80	0.77, 0.83	0.83	0.80, 0.86	0.83	0.80, 0.86	0.81	0.78, 0.84
Sweep 6 (age 42)										
Participated in NCDS V										
No	NA	NA	1.18	1.11, 1.25	1.33	1.24, 1.43	1.28	1.18, 1.39	1.35	1.25, 1.45
Yes	NA	NA	1.00	(reference)	1.00	(reference)	1.00	(reference)	1.00	(reference)
Intends to move in near future										
No	NA	NA	1.00	(reference)	1.00	(reference)	NS	NS	NS	NS
Yes	NA	NA	1.15	1.11, 1.21	1.19	1.12, 1.26	NS	NS	NS	NS
Membership in organisations										
No	NA	NA	NS	NS	1.14	1.06, 1.23	1.14	1.06, 1.22	1.14	1.06, 1.23
Yes	NA	NA	NS	NS	1.00	(reference)	1.00	(reference)	1.00	(reference)
Biomedical sweep (age 44)										
Sweep 7 (age 46)										
Marital status - de facto										
Married	NA	NA	NA	NA	NA	NA	NS	NS	1.00	(reference)
Cohabiting (living as a couple)	NA	NA	NA	NA	NA	NA	NS	NS	0.99	0.89, 1.11
Single (and never married)	NA	NA	NA	NA	NA	NA	NS	NS	1.18	1.07, 1.32
Separated, divorced or widowed	NA	NA	NA	NA	NA	NA	NS	NS	1.23	1.12, 1.35

Sweep 8 (age 50)										
Total number of natural children [per child]	NA	NA	NA	NA	NA	NA	NA	NA	1.05	1.03, 1.08
Employer provided pension scheme										
No	NA	NA	NA	NA	NA	NA	NA	NA	1.13	1.06, 1.20
Yes	NA	NA	NA	NA	NA	NA	NA	NA	1.00	(reference)

NA: Not applicable (predictor variable observed concurrently with or subsequent to non-response variable); NS: Not selected. Note that no biomedical sweep variables were selected as consistent predictors of non-response.

FIGURES

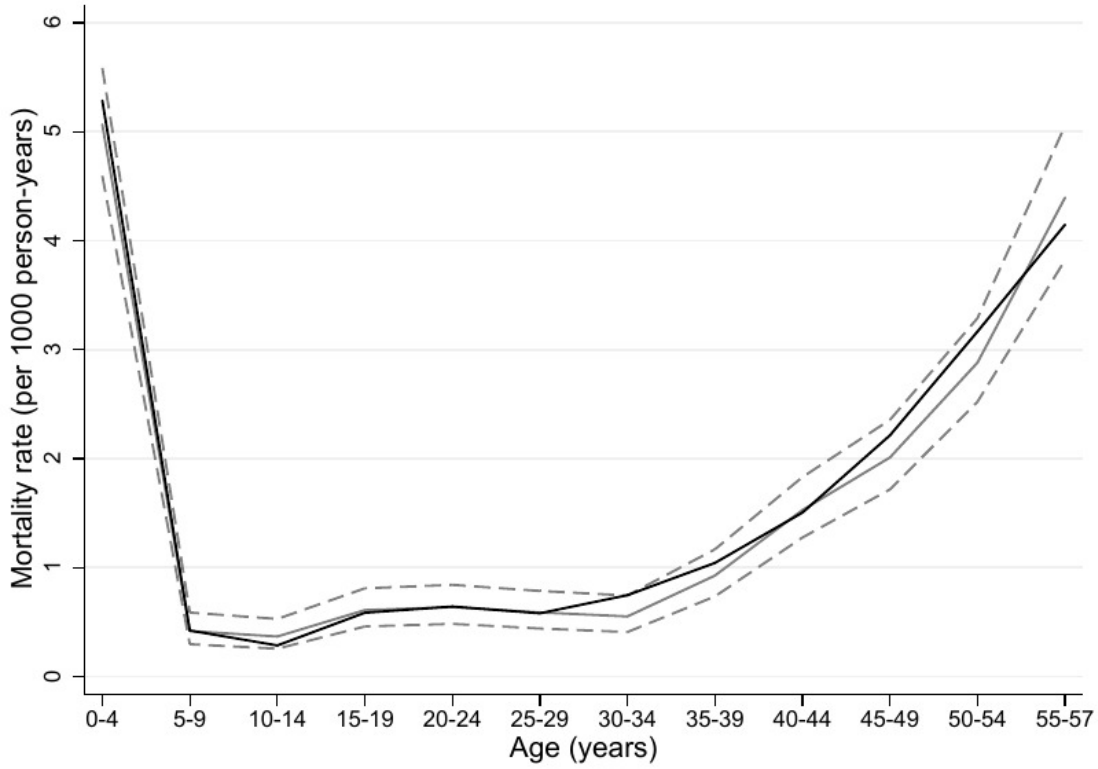


Figure 1. 1958 British National Child Development Study (England and Wales sample; grey solid line (estimate) and grey dashed lines (95% confidence intervals)) & Office for National Statistics standardised mortality rate for England and Wales (black line).

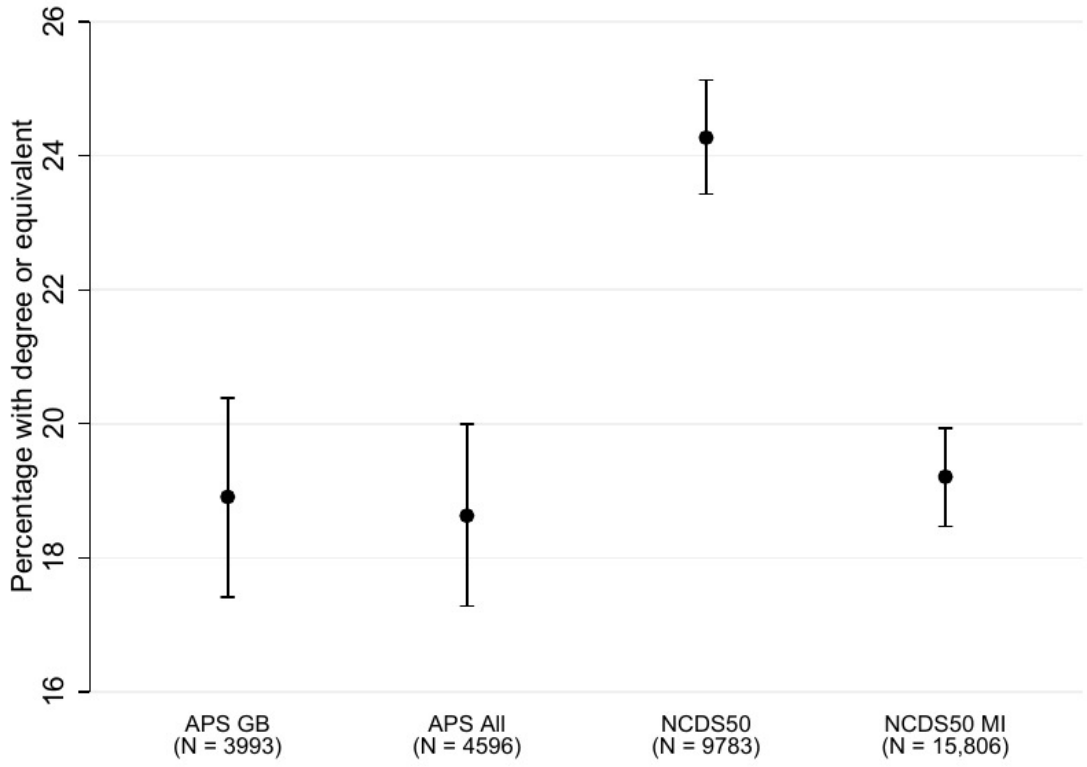


Figure 2. Percentage of those with degree or equivalent at age 50 in the Annual Population Survey and 1958 British National Child Development Study before and after adjustment for missing data. APS GB: Annual Population Survey = Born in Great Britain in 1958 (derived by the Office for National Statistics); APS All: Annual Population Survey - Born in Great Britain or elsewhere in 1958 (derived by the Office for National Statistics); NCDS50: Estimate using observed educational attainment at age 50; NCDS50 MI: Estimate after multiple imputation using predictors of educational attainment at age 50 (see below) and predictors of non-response at age 5 (see Table S10) as auxiliary variables. Predictors of educational attainment at age 50: Maternal interest in cohort member's education at age 7; Overcrowding at age 11; Being off school > 1 month at age 11; Family financial difficulties at age 11; Housing tenure at age 7; Mother reading to CM at age 7; Maternal smoking during pregnancy; Maternal employment (birth to 5 years); Training courses by age 23; Child's positive activities at school age 11; Parity at birth; Nocturnal enuresis at 7; Ever breastfed; Smoking at age 42.

Web Appendix

Contents

Methods S1. Predictors of non-response.

NCDS datasets from the sweeps up to age 50 deposited in the UK Data Service include a total of 17,412 variables that could potentially be used as predictors of non-response. However, many of these variables are so called “routed”, where only cohort members that gave a specific response to a previous question are asked these subsequent questions. For example, variables with information on the presence of specific chronic illnesses are routed on a previous question about the presence of any chronic illness and only those with a chronic illness respond to the subsequent questions. To avoid sample selection the majority of “routed” variables were excluded from the analysis. Exceptions included variables related to occupational social class and employment status. We also excluded binary variables with prevalence less than 1% and variables with item non-response > 50%. We did so as low prevalent categories in binary variables that cannot be collapsed with others would be problematic in the multivariable regression models we employ for variable selection. Similarly, variables with >50% of item non-response in addition to unit non-response would, in combination with missingness in the other predictors of non-response, reduce the available data to <10% in later sweeps. Summary scores were calculated for all scales, further reducing the number of eligible variables. In sweeps where more than one scale was available that taps into the same construct we included in the analysis the one available in most sweeps. Finally, variables that reflect questions used to derive summary measures such as household income, employment status and educational qualifications were not selected as summaries were available.

Methods S2. Derivation of consistent predictors of non-response

Cognitive ability at 7: Principal Component Analysis (PCA) score. PCA indicators were the Problem Arithmetic Test score, Total score on Copying Designs Test, Drawing a Man Test score and the Southgate Group Reading Test score.

Cognitive ability at 11: A general ability test score consisting of 40 verbal and 40 non-verbal items (range 0 to 80). Children were tested individually by teachers, who recorded the answers for the tests. For the verbal items, children were presented with an example set of four words that were linked either logically, semantically, or phonologically. For the non-verbal tasks, shapes or symbols were used. The children were then given another set of three words or shapes or symbols with a blank. Participants were required to select the missing item from a list of five alternatives.

Conduct problems at 11 and 16: Conduct problems and affective symptoms in childhood and adolescence were assessed using the modified version of the Rutter 'A' scale [1]. This version of the scale was completed by the mothers of the participants at ages 7 and 11 years, and from both mother and teachers at age 16. Mother and teacher reports were employed to capture symptoms both at home and school, as is well known that maternal and teacher reports are weakly correlated and that triangulating information from multiple informants may bring unique insights into children's behaviour and may predict poor child and adolescent outcomes in ways that the individual informants' reports do not [2]. Conduct problems refer to behaviour such as being disobedient, destructive, being irritable and being involved in fights. A latent summary score of four conduct problems derived from a 2 parameter was included in the analysis. We derived latent summary of conduct problems at 16 by modelling the probability of response to the Rutter items with a 2 parameter probit latent variable measurement model [3, 4] and calculated a latent trait summary score.

Social participation at age 23: Sum of voluntary activities.

Social Capital at age 33: Number of people you turn to for support.

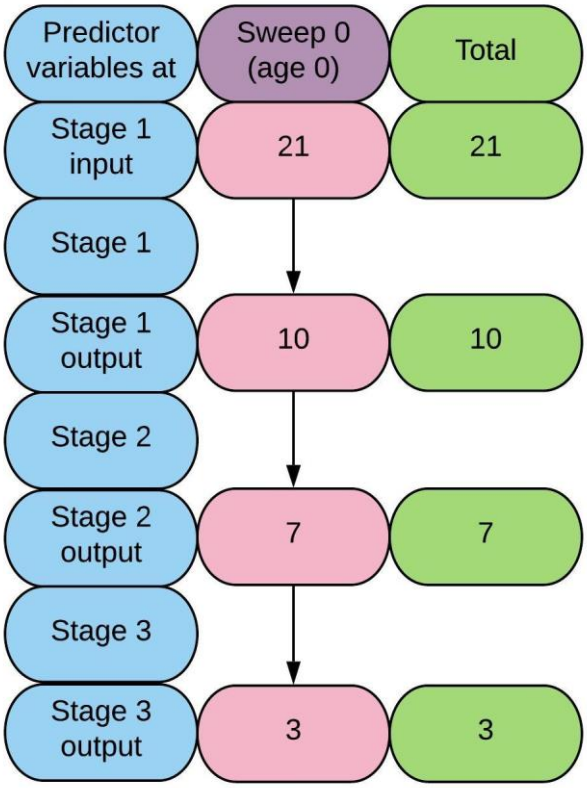
Social participation at age 42: Ever being a member of an organisation (political party, environmental charity, voluntary group, women groups, parents/school/tenant organisations).

Social participation at age 50: Sum score of membership in various organisations: Political party, Trade Union, Environmental group, Parents, School association, Residents Group, neighbourhood watch, Religious Group or Church Organisation, Voluntary Service group, Other Community, civic group, Social, Working men's club, Sports club, Professional organisation, Scouts, Guides organisation, Other Organisation.

References

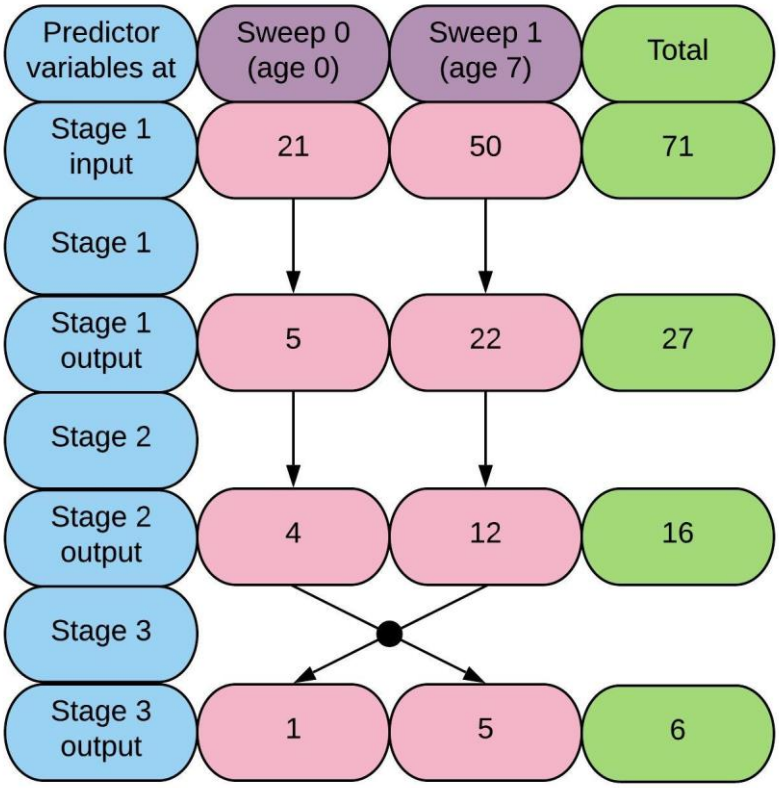
1. Rutter, M., J. Tizard, and K. Whitmore, *Education, health and behaviour*. 1970: Longman Publishing Group.
2. De Los Reyes, A., *Introduction to the special section: More than measurement error: Discovering meaning behind informant discrepancies in clinical assessments of children and adolescents.* J Clin Child Adolesc Psychol, 2011. **40**(1): p. 1-9.
3. Muthén, B., *A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators.* Psychometrika, 1984. **49**(1): p. 115-132.
4. Rabe-Hesketh, S. and A. Skrondal, *Classical latent variable models for medical research.* Statistical Methods in Medical Research, 2008. **17**(1): p. 5-32.

Figure S1. Predictors of non-response at sweep 1 (age 7).



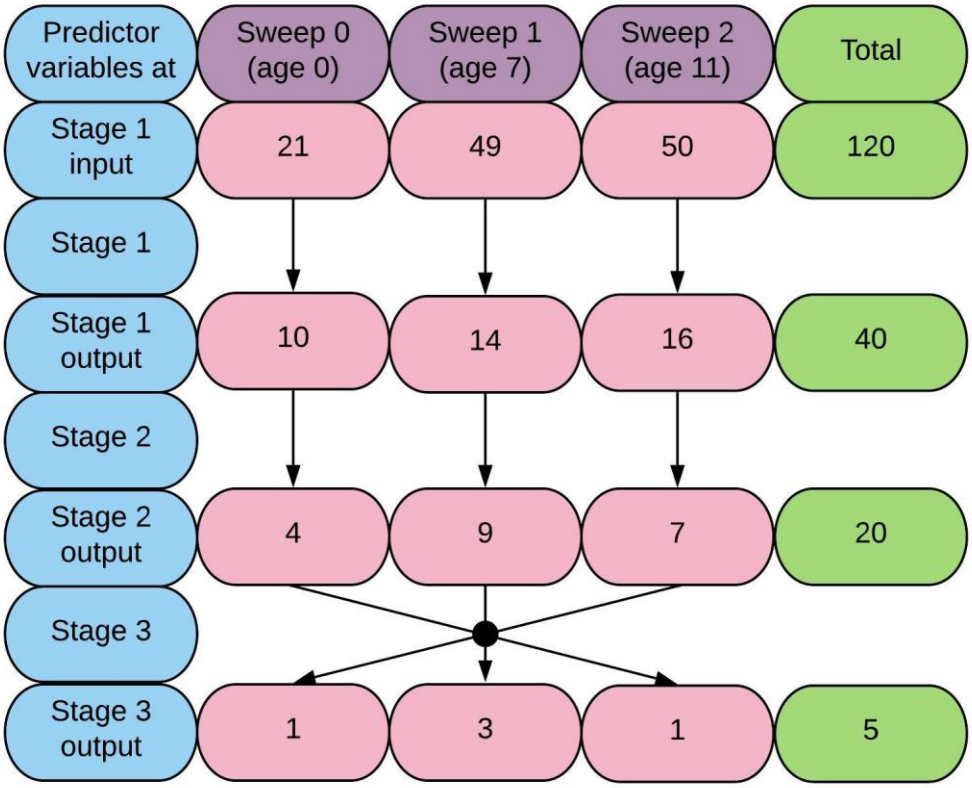
At sweep 1 (age 7) there were 21 eligible predictor variables from sweep 0 (Stage 1 input). Of these, 10 variables were associated with non-response at sweep 1 in univariable models (Stage 1 output). After competing within sweep in multivariable models to predict non-response at sweep 1, 7 variables were retained (Stage 2 output). After competing across all sweeps for the prediction of non-response at sweep 1, 3 variables were retained (Stage 3 output).

Figure S2. Predictors of non-response at sweep 2 (age 11).



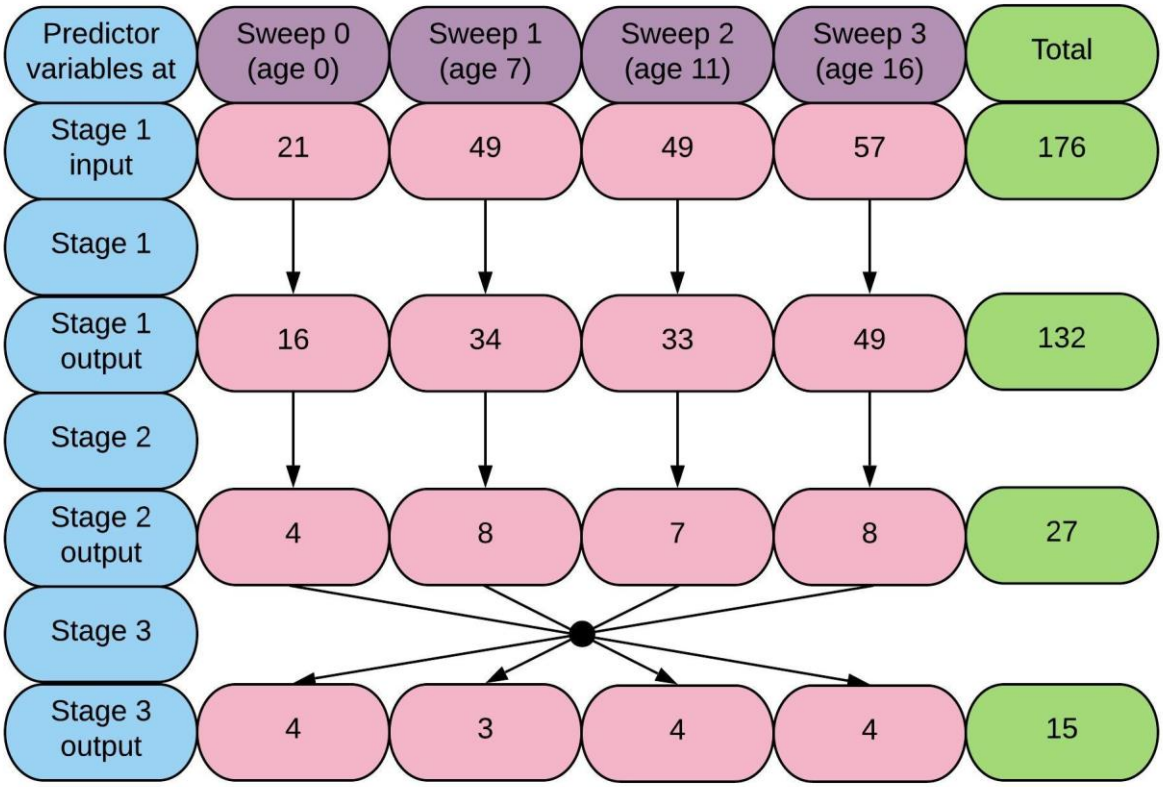
At sweep 2 (age 11) there were 71 eligible predictor variables across sweeps 0 to 1 (Stage 1 input). Of these, 27 variables were associated with non-response at sweep 2 in univariable models (Stage 1 output). After competing within sweep in multivariable models to predict non-response at sweep 2, 16 variables were retained (Stage 2 output). After competing across all sweeps for the prediction of non-response at sweep 2, 6 variables were retained (Stage 3 output).

Figure S3. Predictors of non-response at sweep 3 (age 16).



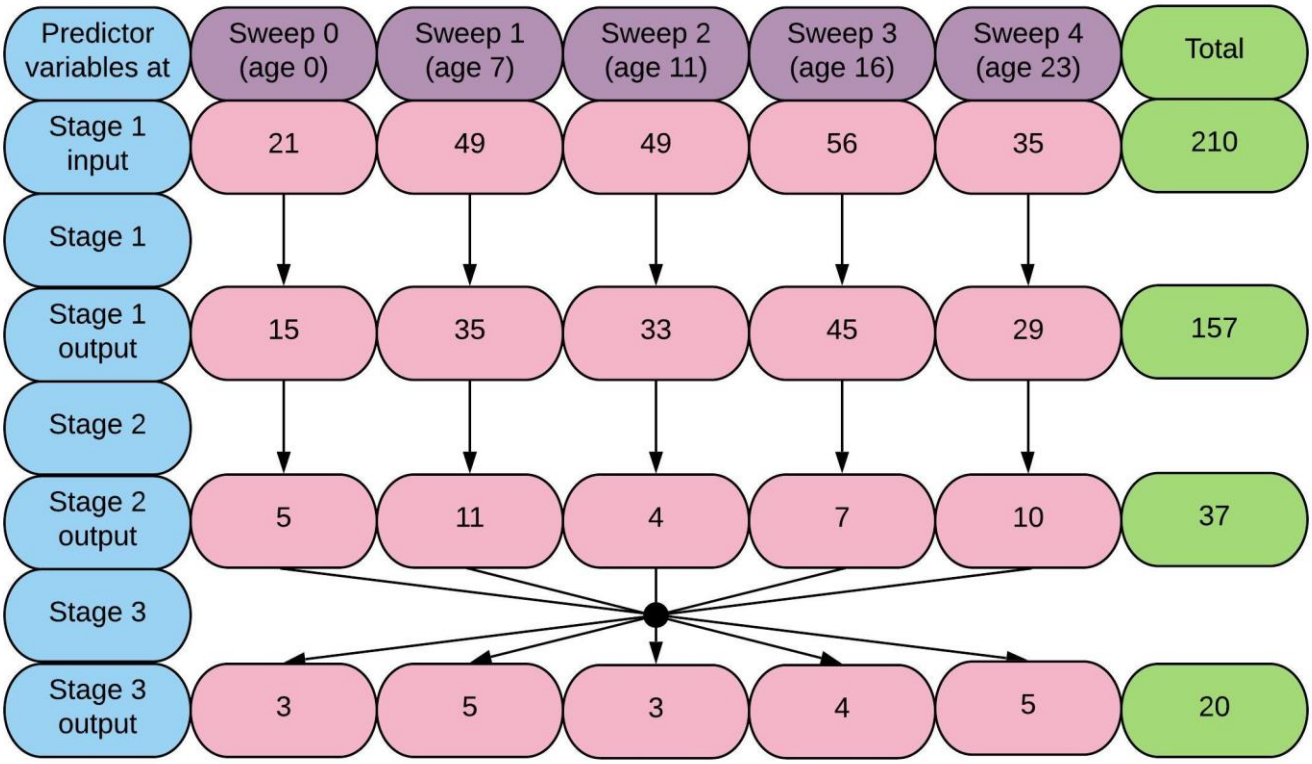
At sweep 3 (age 16) there were 120 eligible predictor variables across sweeps 0 to 2 (Stage 1 input). Of these, 40 variables were associated with non-response at sweep 3 in univariable models (Stage 1 output). After competing within sweep in multivariable models to predict non-response at sweep 3, 20 variables were retained (Stage 2 output). After competing across all sweeps for the prediction of non-response at sweep 3, 5 variables were retained (Stage 3 output).

Figure S4. Predictors of non-response at sweep 4 (age 23).



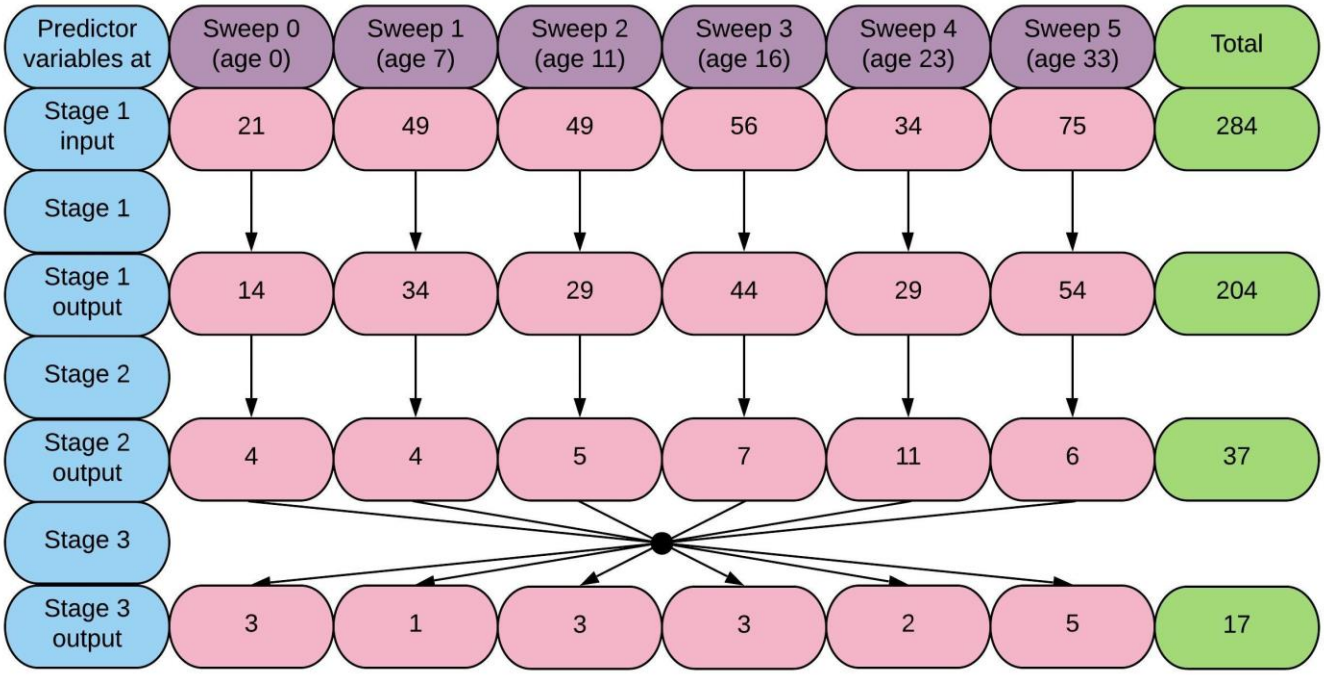
At sweep 4 (age 23) there were 176 eligible predictor variables across sweeps 0 to 3 (Stage 1 input). Of these, 132 variables were associated with non-response at sweep 4 in univariable models (Stage 1 output). After competing within sweep in multivariable models to predict non-response at sweep 4, 27 variables were retained (Stage 2 output). After competing across all sweeps for the prediction of non-response at sweep 4, 15 variables were retained (Stage 3 output).

Figure S5. Predictors of non-response at sweep 5 (age 33).



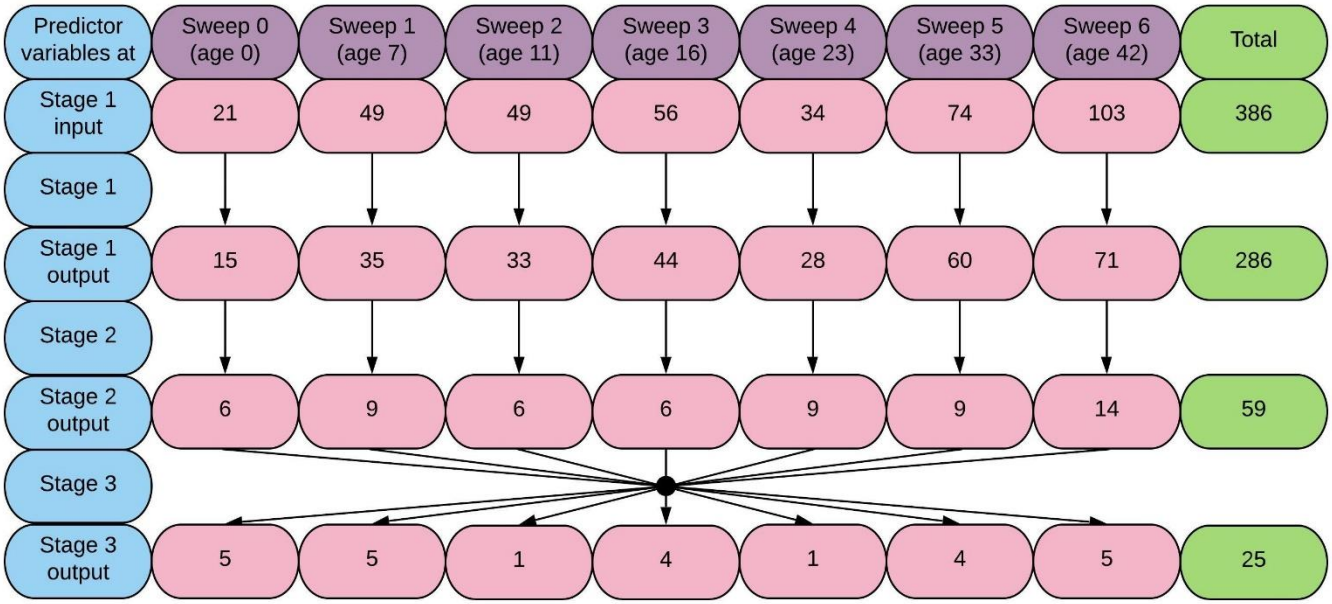
At sweep 5 (age 33) there were 210 eligible predictor variables across sweeps 0 to 4 (Stage 1 input). Of these, 157 variables were associated with non-response at sweep 5 in univariable models (Stage 1 output). After competing within sweep in multivariable models to predict non-response at sweep 5, 37 variables were retained (Stage 2 output). After competing across all sweeps for the prediction of non-response at sweep 5, 20 variables were retained (Stage 3 output).

Figure S6. Predictors of non-response at sweep 6 (age 42).



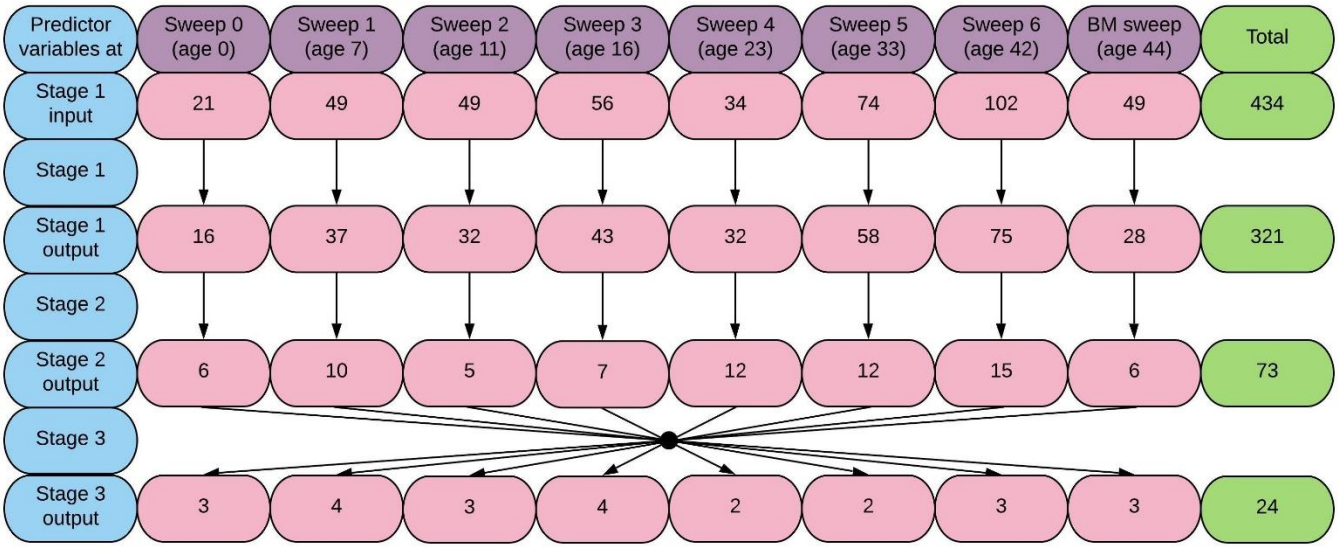
At sweep 6 (age 42) there were 284 eligible predictor variables across sweeps 0 to 5 (Stage 1 input). Of these, 204 variables were associated with non-response at sweep 6 in univariable models (Stage 1 output). After competing within sweep in multivariable models to predict non-response at sweep 6, 37 variables were retained (Stage 2 output). After competing across all sweeps for the prediction of non-response at sweep 6, 17 variables were retained (Stage 3 output).

Figure S7. Predictors of non-response at the biomedical sweep (age 44).



At the biomedical sweep (age 44) there were 386 eligible predictor variables across sweeps 0 to 6 (Stage 1 input). Of these, 286 variables were associated with non-response at the biomedical sweep in univariable models (Stage 1 output). After competing within sweep in multivariable models to predict non-response at the biomedical sweep, 59 variables were retained (Stage 2 output). After competing across all sweeps for the prediction of non-response at the biomedical sweep, 25 variables were retained (Stage 3 output).

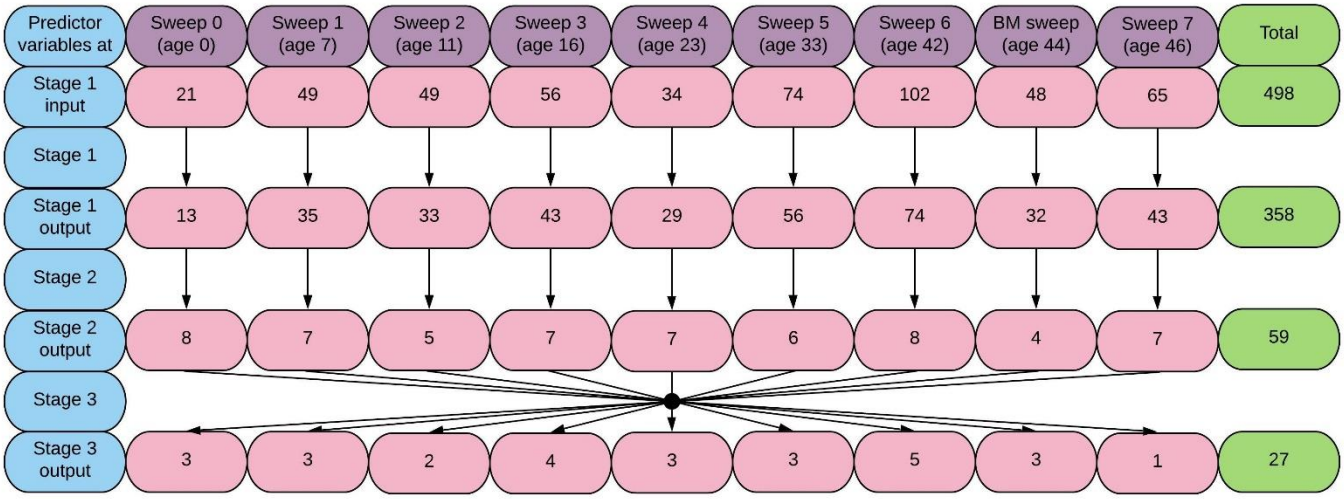
Figure S8. Predictors of non-response at sweep 7 (age 46).



BM: Biomedical.

At sweep 7 (age 46) there were 434 eligible predictor variables across sweeps 0 to the biomedical sweep (Stage 1 input). Of these, 321 variables were associated with non-response at sweep 7 in univariable models (Stage 1 output). After competing within sweep in multivariable models to predict non-response at sweep 7, 73 variables were retained (Stage 2 output). After competing across all sweeps for the prediction of non-response at sweep 7, 24 variables were retained (Stage 3 output).

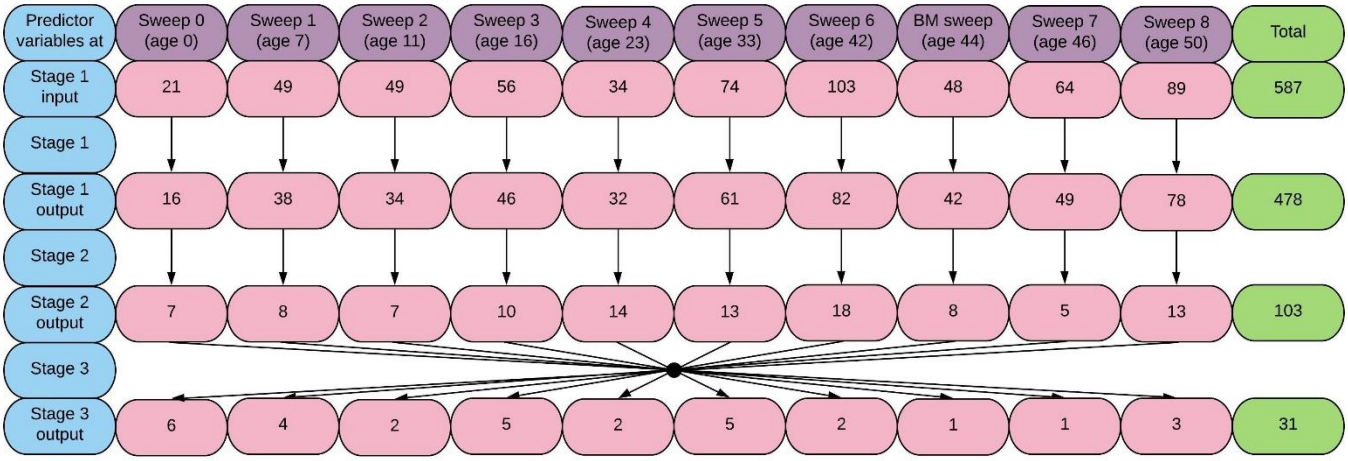
Figure S9. Predictors of non-response at sweep 8 (age 50).



BM: Biomedical.

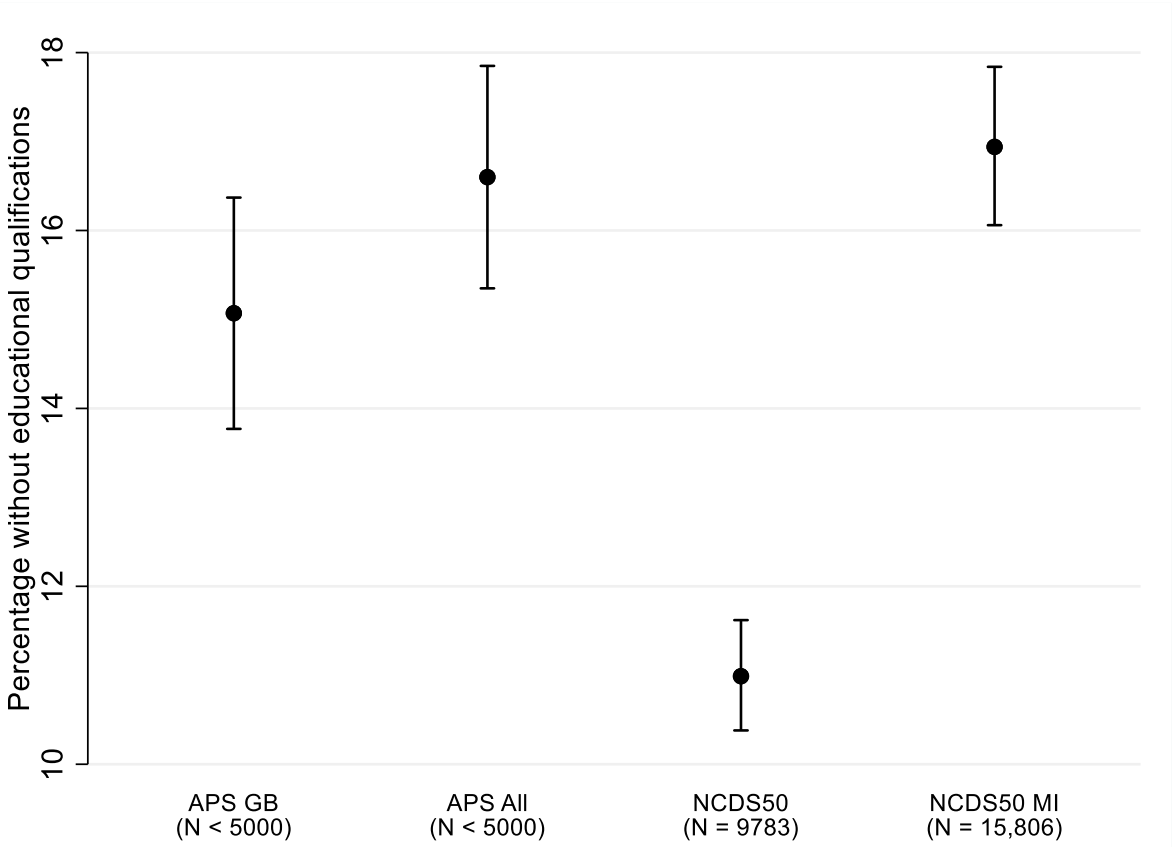
At sweep 8 (age 50) there were 498 eligible predictor variables across sweeps 0 to 7 (Stage 1 input). Of these, 358 variables were associated with non-response at sweep 8 in univariable models (Stage 1 output). After competing within sweep in multivariable models to predict non-response at sweep 8, 59 variables were retained (Stage 2 output). After competing across all sweeps for the prediction of non-response at sweep 8, 27 variables were retained (Stage 3 output).

Figure S10. Predictors of non-response at sweep 9 (age 55).



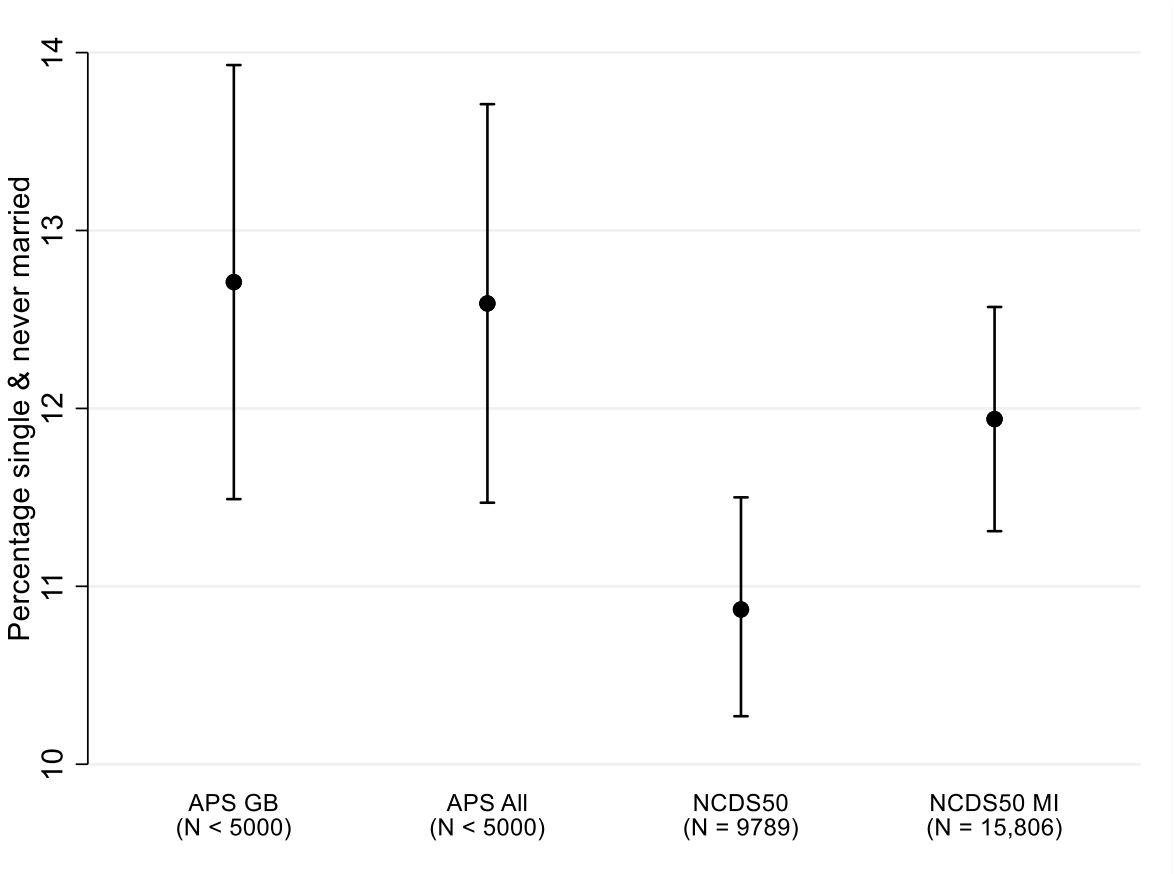
BM: Biomedical.
At sweep 9 (age 55) there were 587 eligible predictor variables across sweeps 0 to 8 (Stage 1 input). Of these, 478 variables were associated with non-response at sweep 9 in univariable models (Stage 1 output). After competing within sweep in multivariable models to predict non-response at sweep 9, 103 variables were retained (Stage 2 output). After competing across all sweeps for the prediction of non-response at sweep 9, 31 variables were retained (Stage 3 output).

Figure S11. Percentage of those without educational qualifications at age 50 in the Annual Population Survey and NCDS before and after adjustment for missing data.



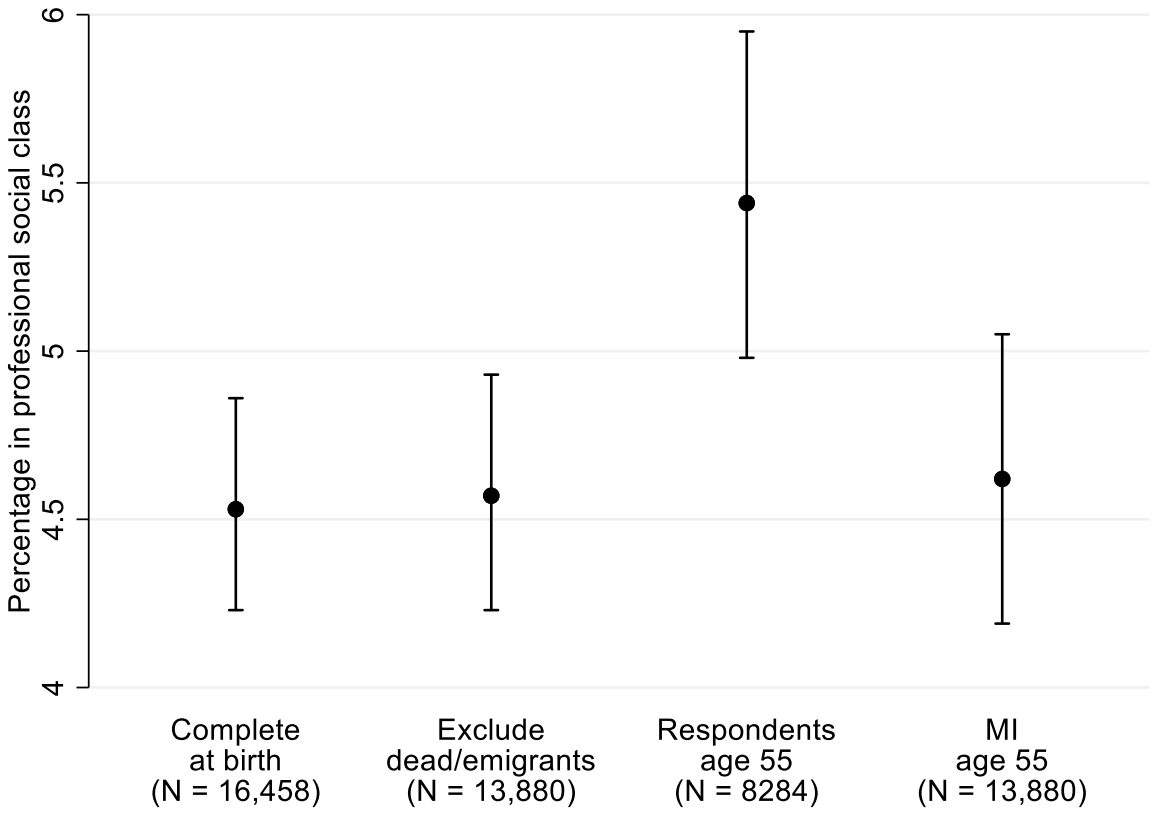
APS GB: Annual Population Survey = Born in Great Britain in 1958 (derived by the Office for National Statistics, N = 3993)
APS All: Annual Population Survey - Born in Great Britain or elsewhere in 1958 (derived by the Office for National Statistics; N = 4596)
NCDS50: Estimate using observed educational attainment at age 50.
NCDS50 MI: Estimate after multiple imputation using predictors of educational attainment at age 50 (see below) and predictors of non-response at age 50 (see Table S10) as auxiliary variables.
Predictors of educational attainment at age 50: Maternal interest in cohort member's education at age 7; Overcrowding at age 11; Being off school > 1 month at age 11; Family financial difficulties at age 11; Housing tenure at age 7; Mother reading to CM at age 7; Maternal smoking during pregnancy; Maternal employment (birth to 5 years); Training courses by age 23; Child's positive activities at school age 11; Parity at birth; Nocturnal enuresis at 7; Ever breastfed; Smoking.

Figure S12. Percentage of those single and never married by age 50 in the Annual Population Survey and NCDS before and after adjustment for missing data.



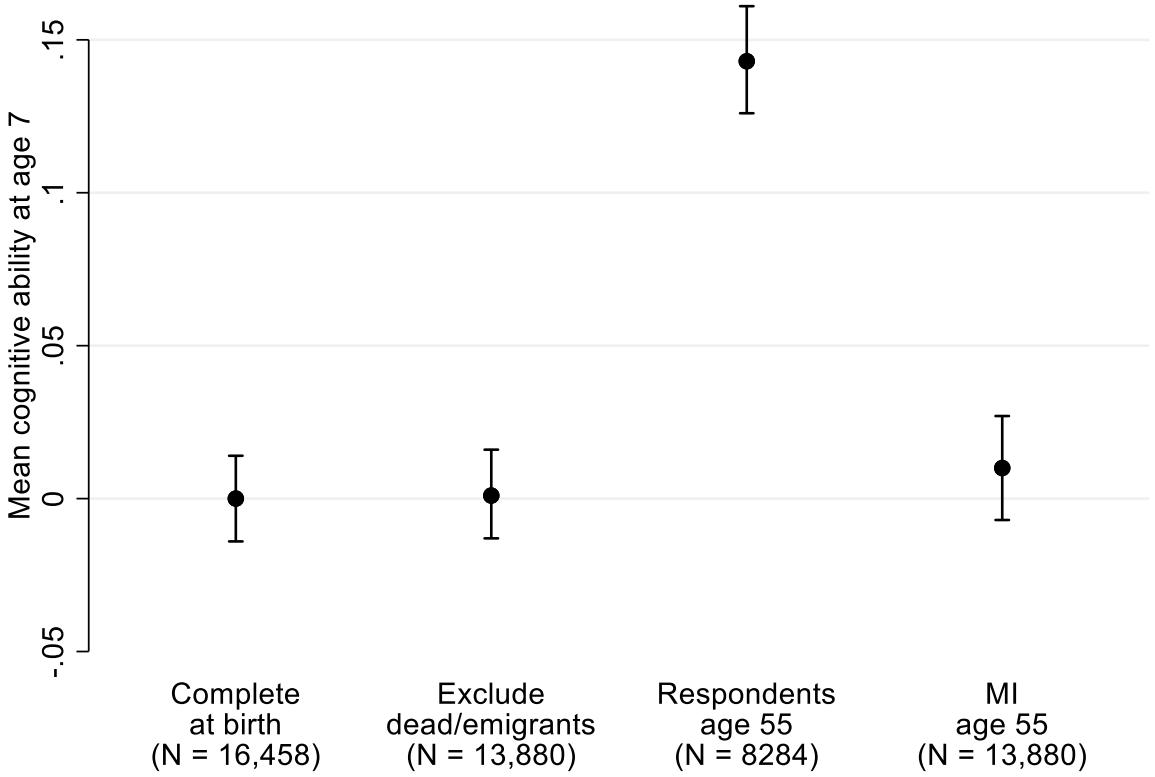
APS GB: Annual Population Survey = Born in Great Britain in 1958 (derived by the Office for National Statistics, N = 3993)
APS All: Annual Population Survey - Born in Great Britain or elsewhere in 1958 (derived by the Office for National Statistics; N = 4596)
NCDS50: Estimate using observed marital status at age 50.
NCDS50 MI: Estimate after multiple imputation using predictors of marital status at age 50 (see below) and predictors of non-response at age 50 (see Table S10) as auxiliary variables.
Predictors of marital status at age 50: Marital status at ages 23, 33, 42, 44 and 46.

Figure S13. Social class of mother's husband at birth before and after adjustment for missing data.



Imputation phase of MI included all predictors of response at age 55 (see Table S11) and social class at birth only for cohort members that participated at age 55.

Figure S14. Cognitive ability at age 7 before and after adjustment for missing data.



Imputation phase of MI included all predictors of response at age 55 (see Table S11) and cognitive ability at age 7 only for cohort members that participated at age 55.

Table S1. 1958 British National Child Development Study survey response by sweep.

	Sweep 0 (age 0)	Sweep 1 (age 7)	Sweep 2 (age 11)	Sweep 3 (age 16)	Sweep 4 (age 23)	Sweep 5 (age 33)	Sweep 6 (age 42)	Biomedical sweep (age 44)	Sweep 7 (age 46)	Sweep 8 (age 50)	Sweep 9 (age 55)
Productive	17415	15425	15337	14654	12537	11469	11419	9377	9534	9790	9137
Refusal	0	80	797	1151	915	1365	1148	2829	1448	1214	582
Non-contact	218	1036	406	786	1675	1394	1832	792	612	835	860
Other unproductive	0	173	202	295	413	953	263	31	109	332	491
Ineligible	0	0	0	0	0	0	13	65	11	81	0
Not Issued*	925	548	275	0	862	993	1415	2908	4248	3553	4543
Not Issued – Emigrant	0	475	701	799	1196	1335	1268	1234	1272	1293	1286
Not Issued – Dead	0	821	840	873	960	1049	1200	1322	1324	1460	1659
Total	18558	18558	18558	18558	18558	18558	18558	18558	18558	18558	18558

*Sweep 0-2: Immigrant – not resident in Great Britain; Sweep 4-9: no address or refusal to participate.

Information taken from Johnson J, Brown M. National Child Development Study: User Guide to the Response and Deaths Datasets. London: Centre for Longitudinal Studies. 2015.

Table S2. Age-specific mortality rates – 1958 British National Child Development Study (NCDS) vs Office for National Statistics (ONS) data (England and Wales).

	NCDS					*ONS*
			Rate (per 1000 person-years)			
Age group	*Deaths*	*Person-years*	*Estimate*	*95% CI*		*Rate (per 1000 person-years)*
0-4	403	79544	5.066	4.595	5.586	5.286
5-9	33	78958	0.418	0.297	0.588	0.423
10-14	29	78823	0.368	0.256	0.529	0.286
15-19	48	78638	0.610	0.460	0.810	0.586
20-24	50	78382	0.638	0.484	0.842	0.643
25-29	46	78158	0.589	0.441	0.786	0.581
30-34	43	77922	0.552	0.409	0.744	0.747
35-39	72	77671	0.927	0.736	1.168	1.043
40-44	118	77205	1.528	1.276	1.831	1.505
45-49	154	76562	2.011	1.718	2.356	2.214
50-55	218	75682	2.881	2.522	3.289	3.166
55-57	197	44825	4.395	3.822	5.054	4.147

ONS rate: population estimates from the Human Mortality Database.

Table S3. Number of cohort members contributing to each fitted stage 2 model.

	Non-response									
Predictors	Sweep 1 (age 7)	Sweep 2 (age 11)	Sweep 3 (age 16)	Sweep 4 (age 23)	Sweep 5 (age 33)	Sweep 6 (age 42)	Biomedical sweep (age 44)	Sweep 7 (age 46)	Sweep 8 (age 50)	Sweep 9 (age 55)
Sweep 0 (age 0)	11,571	15,898	12,827	9,266	8,863	8,816	8,875	8,505	9,947	8,674
Sweep 1 (age 7)	NA	7,878	11,310	7,847	7,834	7,812	7,003	7,032	7,716	6,796
Sweep 2 (age 11)	NA	NA	10,893	6,717	6,962	7,100	6,621	7,032	6,970	6,819
Sweep 3 (age 16)	NA	NA	NA	4,685	5,039	5,020	5,020	5,034	4,940	4,671
Sweep 4 (age 23)	NA	NA	NA	NA	8,908	8,108	8,045	7,985	7,977	7,760
Sweep 5 (age 33)	NA	NA	NA	NA	NA	6,132	4,854	5,399	4,904	4,835
Sweep 6 (age 42)	NA	NA	NA	NA	NA	NA	5,790	7,136	6,299	6,677
BM sweep (age 44)	NA	NA	NA	NA	NA	NA	NA	5,353	4,763	3,543
Sweep 7 (age 46)	NA	NA	NA	NA	NA	NA	NA	NA	6,703	6,595
Sweep 8 (age 50)	NA	NA	NA	NA	NA	NA	NA	NA	NA	3,670

Table S4. Estimated risk ratios and 95% confidence intervals for predictors of non-response at sweep 1 (age 7) (n = 17,262).

Sweep	Variable	RR	95% CI
Sweep 0	Region		
(age 0)	North	1.12	0.85, 1.49
	Midlands	1.23	0.91, 1.68
	East & South East	1.59	1.20, 2.12
	South & South West	1.48	1.09, 2.02
	Wales	1.00	(reference)
	Scotland	1.35	0.99, 1.84
	Number of persons per room [per person]	1.10	1.05, 1.16
	Social class of mother's husband		
	I	1.00	(reference)
	II	0.66	0.51, 0.84
	III non-manual	0.65	0.49, 0.86
	III manual	0.59	0.47, 0.73
	IV	0.72	0.57, 0.92
	V	0.80	0.62, 1.02

Results from sequential multiple imputation analyses in which potential predictors of non-response at a given sweep are adjusted for previously identified potential predictors of non-response at that sweep and previous sweeps (i.e. not at subsequent sweeps).

Table S5. Estimated risk ratios and 95% confidence intervals for predictors of non-response at sweep 2 (age 11) (n = 17,017).

Sweep	Variable	RR	95% CI
Sweep 0 (age 0)	Mother's present marital status		
	Married/Twice married	1.00	(reference)
	Unmarried/Stable union/Separated, divorced, widowed	1.65	1.35, 2.01
Sweep 1 (age 7)	Number of kids under 21 in the household, including living away [per kid]	0.91	0.87, 0.95
	Common difficulties age 7 (mother) [per difficulty]	0.90	0.86, 0.94
	Hospital admissions [per admission]	0.91	0.86, 0.96
	Cognitive ability summary [per unit]	0.85	0.80, 0.91
	Non-response at sweep 1		
	Respondent	5.76	5.28, 6.28
	Non-respondent	1.00	(reference)

Results from sequential multiple imputation analyses in which potential predictors of non-response at a given sweep are adjusted for previously identified potential predictors of non-response at that sweep and previous sweeps (i.e. not at subsequent sweeps).

Table S6. Estimated risk ratios and 95% confidence intervals for predictors of non-response at sweep 3 (age 16) (n = 16,886).

Sweep	Variable	RR	95% CI
Sweep 0 (age 0)	Region		
	North	1.17	0.94, 1.45
	Midlands	1.39	1.11, 1.73
	East & South East	1.70	1.38, 2.10
	South & South West	1.25	0.99, 1.58
	Wales	1.00	(reference)
	Scotland	0.94	0.73, 1.20
Sweep 1 (age 7)	Number of kids under 21 in the household, including living away [per kid]	0.92	0.89, 0.95
	Mother worked birth to 5		
	No	1.20	1.08, 1.33
	Yes	1.00	(reference)
	Ever breastfed		
	Never breastfed	1.21	1.10, 1.35
	Ever breastfed	1.00	(reference)
Sweep 2 (age 11)	Non-response at sweeps 1-2		
	Complete response	1.00	(reference)
	Incomplete response	2.84	2.62, 3.06

Results from sequential multiple imputation analyses in which potential predictors of non-response at a given sweep are adjusted for previously identified potential predictors of non-response at that sweep and previous sweeps (i.e. not at subsequent sweeps).

Table S7. Estimated risk ratios and 95% confidence intervals for predictors of non-response at sweep 4 (age 23) (n = 16,402).

Sweep	Variable	RR	95% CI
Sweep 0 (age 0)	Region		
	North	1.24	1.06, 1.44
	Midlands	1.19	1.02, 1.40
	East & South East	1.45	1.25, 1.69
	South & South West	1.14	0.96, 1.34
	Wales	1.00	(reference)
	Scotland	1.14	0.96, 1.35
	Number of persons per room [per person]	1.11	1.08, 1.14
	Sex of child		
	Male	1.18	1.12, 1.25
	Female	1.00	(reference)
	Social class of mother's husband		
	I	1.00	(reference)
	II	1.01	0.85, 1.21
	III non-manual	0.91	0.75, 1.10
	III manual	1.13	0.96, 1.32
	IV	1.14	0.96, 1.36
	V	1.46	1.23, 1.73
Sweep 1 (age 7)	Family moves since child's birth [per move]	1.10	1.08, 1.12
	Cognitive ability summary [per unit]	0.86	0.83, 0.89
	Dad reads to child		
	Every week sometimes	1.00	(reference)
	Hardly ever	1.13	1.06, 1.22
Sweep 2 (age 11)	Area of world in which mother born		
	British islands	1.00	(reference)
	Eire & Ulster	1.30	1.13, 1.50
	Europe including USSR	1.02	0.83, 1.26
	Outside Europe	1.49	1.29, 1.72
	Number of family moves since child's birth [per move]	1.09	1.05, 1.12
	Cognitive ability summary [per 10 units]	0.91	0.88, 0.94
	Number of household amenities [per unit]	0.91	0.88, 0.95
Sweep 3 (age 16)	Number of family moves since child's birth [per move]	1.07	1.04, 1.11
	Sum of favourable learning environments/outcomes re sex educ etc) [per 10 units]	0.88	0.82, 0.94
	Conduct problems [per unit]	1.10	1.07, 1.13
	Non-response at sweeps 1-3		
	Complete response	1.00	(reference)

	Incomplete response	2.10	1.99, 2.22

Results from sequential multiple imputation analyses in which potential predictors of non-response at a given sweep are adjusted for previously identified potential predictors of non-response at that sweep and previous sweeps (i.e. not at subsequent sweeps).

Table S8. Estimated risk ratios and 95% confidence intervals for predictors of non-response at sweep 5 (age 33) (n = 16,174).

Sweep	Variable	RR	95% CI
Sweep 0 (age 0)	Number of persons per room [per person]	1.11	1.09, 1.13
	Sex of child		
	Male	1.22	1.16, 1.28
	Female	1.00	(reference)
	Social class of mother's husband		
	I	1.00	(reference)
	II	1.06	0.90, 1.24
	III non-manual	1.05	0.89, 1.25
	III manual	1.21	1.04, 1.40
	IV	1.30	1.11, 1.52
	V	1.72	1.47, 2.00
Sweep 1 (age 7)	Family moves since child's birth [per move]	1.04	1.02, 1.06
	Social problems (alcoholism etc.) [per problem]	1.10	1.07, 1.13
	Cognitive ability summary [per unit]	0.87	0.84, 0.89
	Summary of medical conditions [per condition]	0.96	0.94, 0.98
	Ever breastfed		
	Never breastfed	1.11	1.04, 1.17
	Ever breastfed	1.00	(reference)
Sweep 2 (age 11)	Child's positive activities outside school [per 10 activities]	0.89	0.84, 0.94
	Cognitive ability summary [per 10 units]	0.89	0.87, 0.92
	Number of household amenities per unit]	0.93	0.90, 0.97
Sweep 3 (age 16)	Number of family moves since child's birth [per move]	1.06	1.03, 1.08
	How long since child drank alcohol		
	Less than 1 week	1.00	(reference)
	2 to 4 weeks	0.97	0.88, 1.07
	5+ weeks	1.04	0.95, 1.13
	Do not remember	1.11	1.01, 1.22
	Never had one	1.27	1.14, 1.41
	Test 2 – mathematics comprehension [per 10 units]	0.82	0.76, 0.88
	Sum of favourable learning environments/outcomes re sex educ etc) [per 10 units]	0.86	0.81, 0.91
Sweep 4 (age 23)	Type of current accommodation		
	House	1.00	(reference)
	Bungalow	0.92	0.76, 1.11

PB flat	1.23	1.14, 1.33
SC flat	1.13	1.00, 1.27
Other	1.11	0.94, 1.32
Voted in 1979 general election		
Didn't vote	1.24	1.17, 1.32
Voted	1.00	(reference)
Economic status		
Economically inactive	1.10	0.99, 1.21
Full-time education	1.12	0.92, 1.36
Employed	1.00	(reference)
Unemployed	1.20	1.10, 1.31
Number of voluntary activities (youth club, church etc.)	0.94	0.91, 0.97
Non-response at sweeps 1-4		
Complete response	1.00	(reference)
Incomplete response	2.33	2.21, 2.46

Results from sequential multiple imputation analyses in which potential predictors of non-response at a given sweep are adjusted for previously identified potential predictors of non-response at that sweep and previous sweeps (i.e. not at subsequent sweeps).

Table S9. Estimated risk ratios and 95% confidence intervals for predictors of non-response at sweep 6 (age 42) (n = 16,091).

Sweep	Variable	RR	95% CI
Sweep 0 (age 0)	Number of persons per room [per person]	1.11	1.09, 1.13
	Sex of child		
	Male	1.19	1.13, 1.25
	Female	1.00	(reference)
	Social class of mother's husband		
	I	1.00	(reference)
	II	0.94	0.80, 1.11
	III non-manual	1.02	0.86, 1.20
	III manual	1.18	1.02, 1.36
	IV	1.22	1.05, 1.43
	V	1.51	1.30, 1.77
Sweep 1 (age 7)	Cognitive ability summary [per unit]	0.83	0.80, 0.85
Sweep 2 (age 11)	Area of world in which father born		
	British islands	1.00	(reference)
	Eire & Ulster	1.14	0.99, 1.31
	Europe including USSR	1.12	0.94, 1.34
	Outside Europe	1.33	1.17, 1.50
	Child's positive activities outside school [per 10 activities]	0.89	0.85, 0.94
	Cognitive ability summary [per 10 units]	0.88	0.85, 0.90
Sweep 3 (age 16)	How long since child drank alcohol		
	Less than 1 week	1.00	(reference)
	2 to 4 weeks	1.05	0.96, 1.14
	5+ weeks	1.08	1.00, 1.18
	Do not remember	1.11	1.01, 1.23
	Never had one	1.27	1.13, 1.42
	Sum of good activities performed outside school [per activity]	0.97	0.96, 0.98
	Conduct problems [per unit]	1.08	1.05, 1.11
Sweep 4 (age 23)	Legal marital status		
	Single	1.05	0.97, 1.13
	Married	1.00	(reference)
	Separated/divorced/widowed	1.32	1.16, 1.51
	Voted in 1979 general election		
	Didn't vote	1.25	1.18, 1.33
	Voted	1.00	(reference)
Sweep 5	Type of accommodation		

(age 33)	Detached house, etc.	1.00	(reference)
	Semi house/bungalow	0.99	0.87, 1.12
	Terraced house	1.01	0.88, 1.14
	Flat/maisonette/Converted flat, rooms, caravan, miscellaneous	1.26	1.11, 1.44
	Current member of a Trade Union/Staff Association		
	None of those	1.15	1.06, 1.25
	Yes-Trade Union	1.00	(reference)
	Social capital score (people turn to for advice, support) [per 10 units]	0.81	0.77, 0.85
	Life contentment score [per unit]	0.95	0.93, 0.98
	Non-response at sweeps 1-5		
	Complete response	1.00	(reference)
	Incomplete response	3.83	3.57, 4.11

Results from sequential multiple imputation analyses in which potential predictors of non-response at a given sweep are adjusted for previously identified potential predictors of non-response at that sweep and previous sweeps (i.e. not at subsequent sweeps).

Table S10. Estimated risk ratios and 95% confidence intervals for predictors of non-response at biomedical sweep (age 44) (n = 16,003).

Sweep	Variable	RR	95% CI
Sweep 0 (age 0)	Number of persons per room [per person]	1.08	1.07, 1.10
	Abnormality during pregnancy		
	No	1.00	(reference)
	Yes	1.07	1.03, 1.11
	Social class of mother's father when she left school		
	I & II	1.00	(reference)
	III non-manual	0.91	0.81, 1.01
	III manual	1.07	1.01, 1.14
	IV	1.02	0.95, 1.11
	V	1.12	1.04, 1.21
	Sex of child		
	Male	1.07	1.03, 1.11
	Female	1.00	(reference)
	Social class of mother's husband		
	I	1.00	(reference)
	II	1.08	0.95, 1.23
	III non-manual	1.14	1.00, 1.30
	III manual	1.25	1.12, 1.40
	IV	1.32	1.16, 1.49
	V	1.55	1.38, 1.75
Sweep 1 (age 7)	Dad stayed on at school after minimum age		
	No	1.12	1.06, 1.20
	Yes	1.00	(reference)
	Attendance		
	Good attendance	1.00	(reference)
	Frequent short absences	1.17	1.09, 1.26
	Long absences	1.10	1.02, 1.19
	Social problems (alcoholism etc.) [per problem]	1.04	1.02, 1.06
	Cognitive ability summary [per unit]	0.85	0.83, 0.87
	Body mass index [per kg/m^2]	1.02	1.01, 1.04
Sweep 2 (age 11)	Cognitive ability summary [per 10 units]	0.90	0.88, 0.92
Sweep 3 (age 16)	Emotional or behavioural problem		
	No abnormality	1.00	(reference)
	Any condition or handicap	1.23	1.14, 1.32
	How long since child drank alcohol		
	Less than 1 week	1.00	(reference)
	2 to 4 weeks	1.05	0.99, 1.12

	5+ weeks	1.06	1.00, 1.14
	Do not remember	1.14	1.06, 1.22
	Never had one	1.21	1.11, 1.31
	Test 2 – mathematics comprehension [per 10 units]	0.90	0.85, 0.94
	Conduct problems [per unit]	1.06	1.04, 1.08
Sweep 4 (age 23)	Voted in 1979 general election		
	Didn't vote	1.13	1.08, 1.19
	Voted	1.00	(reference)
Sweep 5 (age 33)	Any work related training course since March 1981		
	No	1.12	1.05, 1.19
	Yes	1.00	(reference)
	Number of hospital admissions since March 1981 [per admission]	0.95	0.93, 0.98
	Driven/ridden after drinking alcohol in last 7 days		
	Doesn't drive	1.14	1.07, 1.21
	Yes	0.88	0.80, 0.96
	No	1.00	(reference)
	Social capital score (people turn to for advice, support) [per 10 units]	0.80	0.77, 0.83
Sweep 6 (age 42)	Normally has access to a car or van		
	Yes	1.00	(reference)
	No	1.12	1.04, 1.20
	Doesn't drive	1.13	1.05, 1.22
	Participated in NCDS V		
	No	1.18	1.11, 1.25
	Yes	1.00	(reference)
	Intends to move in near future		
	No	1.00	(reference)
	Yes	1.15	1.11, 1.21
	Has a computer at home		
	No	1.09	1.04, 1.14
	Yes	1.00	(reference)
	Non-response at sweeps 1-6		
	Complete response	1.00	(reference)
	Incomplete response	3.37	3.17, 3.58

Results from sequential multiple imputation analyses in which potential predictors of non-response at a given sweep are adjusted for previously identified potential predictors of non-response at that sweep and previous sweeps (i.e. not at subsequent sweeps).

Table S11. Estimated risk ratios and 95% confidence intervals for predictors of non-response at sweep 7 (age 46) (n = 15,963).

Sweep	Variable	RR	95% CI
Sweep 0 (age 0)	Number of persons per room [per person]	1.08	1.06, 1.10
	Sex of child		
	Male	1.14	1.10, 1.19
	Female	1.00	(reference)
	Social class of mother's husband		
	I	1.00	(reference)
	II	1.09	0.96, 1.25
	III non-manual	1.13	0.99, 1.29
	III manual	1.27	1.13, 1.43
	IV	1.34	1.18, 1.52
	V	1.62	1.43, 1.83
Sweep 1 (age 7)	Dad stayed on at school after minimum age		
	No	1.13	1.06, 1.20
	Yes	1.00	(reference)
	Attendance		
	Good attendance	1.00	(reference)
	Frequent short absences	1.16	1.07, 1.25
	Long absences	1.09	1.01, 1.18
	Social problems (alcoholism etc.) [per problem]	1.03	1.02, 1.05
	Cognitive ability summary [per unit]	0.83	0.81, 0.85
Sweep 2 (age 11)	Source of family income last year		
	Other sources	1.17	1.09, 1.26
	Employment	1.00	(reference)
	Child's positive activities outside school [per 10 activities]	0.93	0.89, 0.97
	Cognitive ability summary [per 10 units]	0.89	0.88, 0.91
Sweep 3 (age 16)	Local Authority & voluntary schools		
	Comprehensive	1.05	1.00, 1.11
	Grammar	1.10	0.99, 1.22
	Secondary modern	1.00	(reference)
	Other	1.23	1.11, 1.37
	Wish could leave school at 15 – study child		
	Yes	1.15	1.09, 1.22
	No	1.00	(reference)
	Uncertain	1.00	0.93, 1.08
	How long since child drank alcohol		
	Less than 1 week	1.00	(reference)
	2 to 4 weeks	1.06	0.99, 1.13
	5+ weeks	1.09	1.02, 1.17
	Do not remember	1.14	1.06, 1.23
	Never had one	1.26	1.17, 1.37
	Test 2 – mathematics comprehension [per 10 units]	0.87	0.82, 0.92
Sweep 4	Number of accidents since 16th birthday [per accident]	1.03	1.01, 1.04

(age 23)	Voted in 1979 general election		
	Didn't vote	1.16	1.11, 1.22
	Voted	1.00	(reference)
Sweep 5	Voted in 1987 general election		
(age 33)	Didn't vote	1.12	1.06, 1.19
	Voted	1.00	(reference)
	Social capital score (people turn to for advice, support) [per 10 units]	0.83	0.80, 0.86
Sweep 6	Participated in NCDS V		
(age 42)	No	1.33	1.24, 1.43
	Yes	1.00	(reference)
	Intends to move in near future		
	No	1.00	(reference)
	Yes	1.19	1.12, 1.26
	Membership in organisations		
	No	1.14	1.06, 1.23
	Yes	1.00	(reference)
BM sweep	Current legal marital status		
(age 44)	Single, never married	1.04	0.92, 1.17
	Married, first and only	1.00	(reference)
	Remarried	1.13	1.02, 1.24
	Separated/divorced/widowed	1.18	1.10, 1.28
	Is current accommodation owned or rented?		
	Other	1.22	1.11, 1.35
	Owner	1.00	(reference)
	Non-response at sweeps 1-biomedical		
	Complete response	1.00	(reference)
	Incomplete response	7.17	6.53, 7.88

Results from sequential multiple imputation analyses in which potential predictors of non-response at a given sweep are adjusted for previously identified potential predictors of non-response at that sweep and previous sweeps (i.e. not at subsequent sweeps). BM: Biomedical.

Table S12. Estimated risk ratios and 95% confidence intervals for predictors of non-response at sweep 8 (age 50) (n = 15,806).

Sweep	Variable	RR	95% CI
Sweep 0 (age 0)	Number of persons per room [per person]	1.07	1.05, 1.09
	Sex of child		
	Male	1.11	1.07, 1.46
	Female	1.00	(reference)
	Social class of mother's husband		
	I	1.00	(reference)
	II	0.98	0.86, 1.12
	III non-manual	1.05	0.92, 1.20
	III manual	1.18	1.05, 1.32
	IV	1.27	1.12, 1.43
	V	1.45	1.28, 1.63
Sweep 1 (age 7)	Social problems (alcoholism etc.) [per problem]	1.07	1.04, 1.09
	Cognitive ability summary [per unit]	0.84	0.82, 0.86
	Summary of medical conditions [per one condition]	0.97	0.96, 0.98
Sweep 2 (age 11)	Cognitive ability summary [per 10 units]	0.90	0.88, 0.92
	Conduct problems [per unit]	1.04	1.02, 1.06
Sweep 3 (age 16)	Child's school attendance [per 10 units]	0.97	0.96, 0.98
	How long since child drank alcohol		
	Less than 1 week	1.00	(reference)
	2 to 4 weeks	1.04	0.97, 1.11
	5+ weeks	1.02	0.95, 1.10
	Do not remember	1.12	1.04, 1.20
	Never had one	1.21	1.10, 1.32
	Test 2 – mathematics comprehension [per 10 units]	0.88	0.83, 0.93
	Conduct problems [per unit]	1.06	1.04, 1.08
Sweep 4 (age 23)	Legal marital status		
	Single	1.04	0.99, 1.10
	Married	1.00	(reference)
	Separated/divorced/widowed	1.21	1.09, 1.34
	Voted in 1979 general election		
	Didn't vote	1.18	1.13, 1.24
	Voted	1.00	(reference)
	Economic status		
	Economically inactive	1.10	1.02, 1.17
	Full-time education	1.14	0.95, 1.37
	Employed	1.00	(reference)
	Unemployed	1.16	1.08, 1.24
Sweep 5 (age 33)	Voted in 1987 general election		
	Didn't vote	1.16	1.10, 1.23
	Voted	1.00	(reference)
	Social capital score (people turn to for advice, support) [per 10 units]	0.83	0.80, 0.86

	Life contentment score [per unit]	0.96	0.95, 0.98
Sweep 6	Frequency of eating biscuits and cakes of all kinds [per category of decreasing consumption]	1.04	1.03, 1.06
(age 42)	Is current accommodation owned or rented?		
	Other	1.19	1.10, 1.28
	Owner	1.00	(reference)
	Participated in NCDS V		
	No	1.28	1.18, 1.39
	Yes	1.00	(reference)
	Ever wanted improve your maths?		
	No	1.13	1.06, 1.21
	Yes	1.00	(reference)
	Membership in organisations		
	No	1.14	1.06, 1.22
	Yes	1.00	(reference)
BM sweep	Consent to access NHS records		
(age 44)	Consent not given	1.54	1.35, 1.75
	Consent given	1.00	(reference)
	How many children do you have living with you aged 18 or less [per child]	0.91	0.86, 0.95
	How many natural (biological) children have you ever had [per child]	1.08	1.04, 1.13
Sweep 7	Non-response at sweeps 1-7		
(age 46)	Complete response	1.00	(reference)
	Incomplete response	6.28	5.71, 6.91

Results from sequential multiple imputation analyses in which potential predictors of non-response at a given sweep are adjusted for previously identified potential predictors of non-response at that sweep and previous sweeps (i.e. not at subsequent sweeps). BM: Biomedical.

Table S13. Estimated risk ratios and 95% confidence intervals for predictors of non-response at sweep 9 (age 55) (n = 15,613).

Sweep	Variable	RR	95% CI
Sweep 0	Mother's age [per 10 years]	0.93	0.89, 0.97
(age 0)	Number of persons per room [per person]	1.06	1.04, 1.08
	Parity [per child]	1.04	1.02, 1.05
	Social class of mother's father when she left school		
	I & II	1.00	(reference)
	III non-manual	0.89	0.79, 0.99
	III manual	1.08	1.01, 1.15
	IV	1.08	1.00, 1.17
	V	1.17	1.09, 1.27
	Sex of child		
	Male	1.13	1.09, 1.18
	Female	1.00	(reference)
	Social class of mother's husband		
	I & II	1.00	(reference)
	III non-manual	1.11	1.01, 1.22
	III manual	1.35	1.26, 1.43
	IV	1.41	1.31, 1.53
	V	1.69	1.57, 1.82
Sweep 1	Dad stayed on at school after minimum age		
(age 7)	No	1.15	1.07, 1.23
	Yes	1.00	(reference)
	Social problems (alcoholism etc.) [per problem]	1.04	1.02, 1.06
	Cognitive ability summary [per unit]	0.82	0.80, 0.84
	Ever breastfed		
	Never breastfed	1.08	1.03, 1.13
	Ever breastfed	1.00	(reference)
Sweep 2	Cognitive ability summary [per 10 units]	0.88	0.86, 0.89
(age 11)	Conduct problems [per unit]	1.03	1.02, 1.05
Sweep 3	Child receiving help at school – backwardness		
(age 16)	No	1.00	(reference)
	Yes	1.13	1.06, 1.20
	Child's school attendance [per 10 units]	0.97	0.96, 0.98
	How long since child drank alcohol		
	Less than 1 week	1.00	(reference)
	2 to 4 weeks	1.03	0.96, 1.10
	5+ weeks	1.04	0.97, 1.11
	Do not remember	1.12	1.04, 1.19
	Never had one	1.22	1.13, 1.31
	Test 2 – mathematics comprehension [per 10 units]	0.86	0.82, 0.90
	Conduct problems [per unit]	1.05	1.03, 1.07
Sweep 4	Legal marital status		
(age 23)	Single	1.12	1.03, 1.21

	Married	1.00	(reference)
	Separated/divorced/widowed	1.24	1.11, 1.38
	Voted in 1979 general election		
	Didn't vote	1.16	1.11, 1.21
	Voted	1.00	(reference)
Sweep 5 (age 33)	Telephone in home		
	No	1.12	1.05, 1.19
	Yes	1.00	(reference)
	How much physical effort in job [per category]	1.05	1.02, 1.07
	Voted in 1987 general election		
	Didn't vote	1.16	1.11, 1.21
	Voted	1.00	(reference)
	Housing tenure		
	Other	1.14	1.08, 1.21
	Owners	1.00	(reference)
	Social capital score (people turn to for advice, support) [per 10 units]	0.81	0.78, 0.84
Sweep 6 (age 42)	Participated in NCDS V		
	No	1.35	1.25, 1.45
	Yes	1.00	(reference)
	Membership in organisations		
	No	1.14	1.06, 1.23
	Yes	1.00	(reference)
BM sweep (age 44)	Self-rated general health [per category of decreasing health]	1.12	1.06, 1.18
Sweep 7 (age 46)	Marital status - de facto		
	Married	1.00	(reference)
	Cohabiting (living as a couple)	0.99	0.89, 1.11
	Single (and never married)	1.18	1.07, 1.32
	Separated, divorced or widowed	1.23	1.12, 1.35
Sweep 8 (age 50)	Total number of natural children [per child]	1.05	1.03, 1.08
	Employer provided pension scheme		
	No	1.13	1.06, 1.20
	Yes	1.00	(reference)
	Non-response at sweeps 1-8		
	Complete response	1.00	(reference)
	Incomplete response	5.93	5.39, 6.54

Results from sequential multiple imputation analyses in which potential predictors of non-response at a given sweep are adjusted for previously identified potential predictors of non-response at that sweep and previous sweeps (i.e. not at subsequent sweeps). BM: Biomedical.

Table S14. Results from sensitivity analysis using LASSO at Stage 2.

Predictors	Non-response	Stage 1 variables	Stage 2 variables	
			Log-binomial	LASSO
Sweep 0 (age 0)	Sweep 1 (age 7)	10	7	9
	Sweep 2 (age 11)	5	4	5
	Sweep 3 (age 16)	10	4	9
	Sweep 4 (age 23)	16	4	13
	Sweep 5 (age 33)	15	5	9
	Sweep 6 (age 42)	14	4	11
	Biomedical sweep (age 44)	15	6	10
	Sweep 7 (age 46)	16	6	12
	Sweep 8 (age 50)	13	8	11
	Sweep 9 (age 55)	16	7	13
Sweep 1 (age 7)	Sweep 2 (age 11)	22	12	20
	Sweep 3 (age 16)	14	9	12
	Sweep 4 (age 23)	34	8	10
	Sweep 5 (age 33)	35	11	27
	Sweep 6 (age 42)	34	4	14
	Biomedical sweep (age 44)	35	9	25
	Sweep 7 (age 46)	37	10	22
	Sweep 8 (age 50)	35	7	22
	Sweep 9 (age 55)	38	8	28

Table S15. Estimated risk ratios and 95% confidence intervals for predictors of non-response at sweep 1 (age 7) (n = 17,262) after LASSO variable selection at Stage 2

Sweep	Variable	RR	95% CI
Sweep 0	Region		
(age 0)	North	1.11	0.84, 1.48
	Midlands	1.24	0.92, 1.68
	East & South East	1.59	1.20, 2.10
	South & South West	1.48	1.08, 2.01
	Wales	1.00	(reference)
	Scotland	1.34	0.99, 1.82
	Number of persons per room [per person]	1.10	1.05, 1.16
	Abnormality during pregnancy		
	No	1.00	(reference)
	Yes	2.23	2.05, 2.43
	Social class of mother's husband		
	I	1.00	(reference)
	II	0.61	0.48, 0.78
	III non-manual	0.65	0.49, 0.85
	III manual	0.58	0.47, 0.72
	IV	0.73	0.58, 0.93
	V	0.78	0.62, 1.00

Results from sequential multiple imputation analyses in which potential predictors of non-response at a given sweep are adjusted for previously identified potential predictors of non-response at that sweep and previous sweeps (i.e. not at subsequent sweeps).

Table S16. Estimated risk ratios and 95% confidence intervals for predictors of non-response at sweep 2 (age 11) (n = 17,017)) after LASSO variable selection at Stage 2.

Sweep	Variable	RR	95% CI
Sweep 0 (age 0)	Mother's present marital status		
	Married/Twice married	1.00	(reference)
	Unmarried/Stable union/Separated, divorced, widowed	1.64	1.33, 2.01
	Abnormality during pregnancy		
	No	1.00	(reference)
	Yes	1.47	1.34, 1.62
Sweep 1 (age 7)	Number of kids under 21 in the household, including living away [per kid]	0.92	0.88, 0.96
	Common difficulties age 7 (mother) [per difficulty]	0.92	0.88, 0.95
	Cognitive ability summary [per unit]	0.87	0.81, 0.92
	Non-response at sweep 1		
	Respondent	5.49	5.02, 6.00
	Non-respondent	1.00	(reference)

Results from sequential multiple imputation analyses in which potential predictors of non-response at a given sweep are adjusted for previously identified potential predictors of non-response at that sweep and previous sweeps (i.e. not at subsequent sweeps).