Using linked Hospital Episode Statistics data to aid the handling of non-response and restore sample representativeness in the 1958 National Child Development Study

CLS working paper number 2023/1

Nasir Rajah[1], Lisa Calderwood[1], Bianca L De Stavola[2], Katie Harron[2], George B Ploubidis[1] and Richard J Silverwood[1]*

1. Centre for Longitudinal Studies, UCL Social Research Institute, 20 Bedford Way, London WC1H 0AL
2. Population, Policy & Practice Research and Teaching Department, UCL Great Ormond Street Institute of Child Health, 30 Guilford Street, London WC1N 1EH

CENTRE FOR LONGITUDINAL STUDIES

Economic and Social Research Council

Contact the authors
Richard Silverwood
UCL Centre for Longitudinal Studies
r.silverwood@ucl.ac.uk

This working paper was first published in February 2023 by the UCL Centre for Longitudinal Studies.
UCL Social Research Institute
University College London
20 Bedford Way
London WC1H 0AL
www.cls.ucl.ac.uk

The UCL Centre for Longitudinal Studies (CLS) is an Economic and Social Research Council (ESRC) Resource Centre based at the UCL Social Research Institute, University College London. It manages four internationally-renowned cohort studies: the 1958 National Child Development Study, the 1970 British Cohort Study, Next Steps, and the Millennium Cohort Study. For more information, visit www.cls.ucl.ac.uk.

This document is available in alternative formats. Please contact the Centre for Longitudinal Studies.
Tel: +44 (0)20 7612 6875
Email: clsfeedback@ucl.ac.uk

Disclaimer

This working paper has not been subject to peer review.

CLS working papers often represent preliminary work and are circulated to encourage discussion. Citation of such a paper should account for its provisional character. A revised version may be available directly from the author.

Any opinions expressed here are those of the author(s) and not those of the UCL Centre for Longitudinal Studies (CLS), the UCL Social Research Institute, University College London, or the Economic and Social Research Council.

Find out more

Email: clsfeedback@ucl.ac.uk
Visit: cls.ucl.ac.uk
Follow: @CLScohorts

How to cite this paper

Rajah, N., Calderwood, L., De Stavola, B.L., Harron, K., Ploubidis, G.B., Silverwood, R.J. (2023) Using linked Hospital Episode Statistics data to aid the handling of non-response and restore sample representativeness in the 1958 National Child Development Study. CLS Working Paper 2023/1. London: UCL Centre for Longitudinal Studies.

Abstract

Background

There is growing interest in whether linked administrative data have the potential to aid analyses subject to missing data in cohort studies.

Methods

Using linked 1958 National Child Development Study (NCDS; British cohort born in 1958, initial n=17,415) and Hospital Episode Statistics (HES) data, we applied a LASSO variable selection approach to identify HES variables which are predictive of non-response at the age 55 sweep of NCDS. We then included these variables as auxiliary variables in multiple imputation (MI) analyses to explore the extent to which they helped restore sample representativeness of the respondents together with the imputed non-respondents in terms of early life variables (father's social class at birth, cognitive ability at age 7) and relative to external population benchmarks (educational qualifications and marital status at age 55).

Results

We identified 10 HES variables that were predictive of non-response at age 55 in NCDS. For example, cohort members who had been treated for adult mental illness were more than 70% more likely to be non-respondents (risk ratio 1.73; 95% confidence interval 1.17, 2.51). Inclusion of these HES variables in MI analyses only helped to restore sample representativeness to a limited extent. Furthermore, there was essentially no additional gain in sample representativeness relative to analyses using only previously identified survey predictors of non-response (i.e. NCDS rather than HES variables).

Conclusions

Inclusion of administrative data variables aided missing data handling in NCDS to a limited extent. However, these findings may not generalise to other analyses or cohorts.

Keywords

Administrative data; Cohort studies; Data linkage; Missing data; Multiple imputation; Representativeness

Introduction

Sample attrition in longitudinal surveys can lead to bias if the remaining respondents are not representative of the survey's target population. Such selective response is likely to be the norm rather than the exception (1, 2), so appropriate handling of missing data due to attrition (or non-response more generally) is imperative.

Recent decades have seen the establishment of a number of principled methods for the handling of missing data, such as multiple imputation (MI) (3), full information maximum likelihood (FIML) (4) and inverse probability weighting (IPW) (5). Typically, application of such methods relies on an assumption of "missingness at random" (MAR). MAR implies that given the observed values, missingness does not depend on unobserved values or, equivalently, that systematic differences between the missing values and the observed values can be explained by observed data (6). Strategies for reducing bias due to non-response may therefore seek to maximise the plausibility of the MAR assumption. This can be achieved by the inclusion of carefully selected auxiliary variables (variables not of direct substantive interest), either in the imputation phase of MI, directly in FIML analysis, or in the derivation of response weights for IPW. Relevant auxiliary variables are those associated with the underlying values of the variable(s) subject to missingness, particularly those also associated with the probability of missingness (3). An important part of analysing data subject to missingness is often therefore the identification of suitable auxiliary variables.

Variables associated with the underlying values of the variable(s) subject to missingness will generally need to be considered on an analysis-specific basis due to the inclusion of different variables in analytic models. However, since the major driver of missingness in longitudinal surveys will generally be wave (as opposed to item) non-response, variables associated with the probability of missingness can be considered more generically by identifying variables predictive of wave non-response. Analysts can then select variables (assumed to be) associated with the underlying values of the variable(s) subject to missingness from the pool of identified predictors of non-response to include as auxiliary variables. Such predictors of non-response can be identified from within the (often vast) pool of variables previously collected as part of the longitudinal survey.

In recent years, many longitudinal surveys have begun to link administrative records (for example, health, education or financial) for their participants with their data collected as part of the survey. Such linked administrative data often contain broader or more detailed information than conventional survey data and may be more complete, since administrative data typically have the benefit of minimal attrition over time (7). There is therefore substantial interest in whether variables derived from linked administrative data may be helpful as auxiliary variables in analyses of survey data subject to missingness.

In this paper we explore this idea using data from the 1958 National Child Development Study (NCDS), a long-running British birth cohort (8), for which linkage to secondary care data from the Hospital Episode Statistics (HES) database is available (9-11). Previous work in NCDS considering only variables collected as part of the study (i.e. not from linked administrative data) found 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 to be consistently associated with non-response (12).. The main aim of this paper is to explore whether administrative data have the potential to enhance approaches to handling missingness data in cohort studies – a question which has received recent interest in relation to NCDS (13, 14).

Methods

Data

1958 National Child Development Study (NCDS)

The NCDS follows the lives of an initial 17,415 people born in Great Britain in a single week of 1958 (8). Since the birth sweep, NCDS cohort members have been followed up 10 times, with the eleventh sweep currently underway with the cohort members now aged 64. The study includes information on cohort members' physical and educational development, economic circumstances, employment, family life, health behaviour, wellbeing, social participation, biological data and attitudes.

Hospital Episode Statistics (HES)

HES is a collection of databases containing details of all admissions (Admitted Patient Care (APC) and Critical Care (CC)), Accident and Emergency (A&E) attendances and Outpatient (OP) appointments at NHS hospitals in England, maintained by NHS Digital (9). Each HES dataset provides detailed information on admission and discharge or appointment dates, diagnoses, procedures, basic patient demographics, and hospital characteristics (15). The period of data availability differs by dataset, from 1997 for APC, from 2007 for A&E, from 2009 for CC and from 2003 for OP.

Linked NCDS-HES data

Linkage between NCDS and all four HES datasets has recently been undertaken, on the basis of consents obtained at NCDS wave 8 (2008, age 50) (10, 11). Matching was conducted using deterministic linkage based on combinations of the participant's name, sex, date of birth and postcode. The flow of data, from the full sample of NCDS cohort members to the linked samples for each HES dataset, is shown in the data flow diagram in Supplementary Fig. S1. Recent analyses suggest the linkage quality of the NCDS-HES data to be high and the linked sample to retain a good level of population representativeness (16).

In this study we restricted our attention to cohort members who were in the wave 9 (2013, age 55) target population (those who were alive and still living in Great Britain at this point). Individuals outside the target population would not have been in the issued sample for the wave 9 follow-up and therefore could not have responded. As our aim was to identify predictors of non-response and not of mortality or emigration, such individuals were excluded rather than being considered as non-respondents. We used linked HES data from the earliest available date until the end of 2012 when NCDS wave 9 took place, to ensure that we only used HES information which pre-dated the point at which response was sought. The impact of these additional criteria on the sample is shown in the data flow diagram in Fig. 1. Annual Population Survey (APS)

The Annual Population Survey (APS) is a large survey administered yearly by the Office for National Statistics (ONS) (17). It contains approximately 320,000 respondents and covers social and economic aspects of individuals' lives. In this study, we used the APS January-

December 2013 survey (18) to derive population estimates for the variables of interest, limiting our analysis to 55-year-olds.

Variables

NCDS

NCDS non-response at age 55 was captured as a binary variable, defined as cohort members 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.

Predictors of age 55 NCDS non-response, listed in Supplementary Table S1, were previously identified using survey data from the 10 preceding sweeps (birth to age 50) of NCDS (12).

To assess how effective the identified HES predictors of NCDS age 55 non-response were at restoring sample representativeness despite selective attrition, we considered representativeness with respect to two NCDS variables observed in early life and two NCDS variables in observed in later life (subsequently collectively referred to as “analysis variables”): father’s social class at birth (binary variable for father being in the professional social class), cognitive ability at age 7 (continuous principal component analysis score derived using the scores from the problem arithmetic test, copying designs test, drawing a man test and Southgate Group Reading Test), educational qualifications at age 55 (binary variable for no educational qualifications), and marital status at age 55 (binary variable for single and never married).

Linked NCDS-HES data

A total of 58 variables to be considered as potential predictors of NCDS non-response at age 55 were derived across the APC, CC and A&E HES datasets. We aimed to derive as many variables as we could using the information available, though intentionally avoided variables with low sample prevalence which would be unlikely to prove useful as auxiliary variables. We therefore derived variables relating to diagnoses and treatments at a high level (e.g. International Classification of Diseases (ICD)-10 chapters) rather than considering more granular coding. The derived variables relate to the numbers of admissions and appointments, missed appointments, investigations undertaken, diagnoses and treatments received (full details in Supplementary Table S2).

APS

For 55-year-olds in APS, we derived the percentage of individuals who were single and had never been married and the percentage of individuals with no educational qualifications using survey information weighted to the mid-2013 population estimate using the weights provided by the ONS (18).

Statistical analysis

HES predictors of NCDS non-response at wave 9 (age 55)

In order to identify which of the 58 derived HES variables were important predictors of non-response at age 55 in NCDS, we employed the Least Absolute Shrinkage and Selection Operator (LASSO) (19). We included all 58 HES variables in a logistic regression model for non-response and used the LASSO lambda value that minimised mean cross-validated error using 10-fold cross-validation.

In a secondary analysis we used a multi-stage P value-based variable selection approach, similar to that employed by Mostafa et al (12), for comparison with the primary approach using the LASSO (see Supplementary Methods S1).

Restoring sample representativeness

We first explored the associations between the analysis variables of interest and the identified HES and survey predictors of non-response. This allowed us to assess whether the HES/survey predictors of non-response were sufficiently well associated with the analysis variables to constitute useful auxiliary variables. Associations were explored using linear or logistic (as appropriate) regression models, with P values from Wald tests of the parameter(s) presented to summarise the strength of evidence for each association.

We undertook a number of different analyses to assess how effective the identified HES predictors of NCDS age 55 non-response were at restoring sample representativeness despite selective attrition. Full details of the analyses are provided in Supplementary Methods S2 but are briefly summarised here.

The first analysis (“Analysis A”) focused on HES linkage consenters who were eligible for linkage (lived in England for at least one wave been wave 6 and wave 9) and who were within the wave 9 (age 55) target population. These individuals are non-missing for all HES variables since we assumed those with no linked HES record truly had no relevant hospital interactions. These analyses considered sample representativeness in terms of variables observed in early life (father’s social class at birth and cognitive ability at age 7). The distribution of each early life variable when using data on all respondents at that point in time (i.e. birth and age 7 respectively) was compared to the distribution among wave 9 respondents only (to assess bias due to non-response at wave 9) and to the distribution when using HES and/or survey predictors of non-response as auxiliary variables in MI analyses (to assess to what extent sample representativeness can be restored using the selected predictors of non-response).

The second analysis (“Analysis B”) focused on all NCDS cohort members (within the wave 9 target population). This includes individuals who did not consent to HES linkage (or who did consent but were ineligible for linkage) and are therefore missing for all HES variables. Analyses related to restoring sample representativeness of early life NCDS variables involved similar comparisons to those outlined for Analysis A. Analyses related to restoring sample representativeness of later life NCDS variables (educational qualifications at age 55 and marital status at age 55) instead compared the distribution among wave 9 respondents

(the only individuals in whom these variables, being collected at wave 9, could be observed) and the distribution when using survey or survey and HES predictors of non-response as auxiliary variables in MI analyses to population benchmarks derived from APS.

In each analysis we utilised MI with chained equations (20), generating 20 imputed datasets. All analyses were conducted in Stata 16 and R 4.0.3.

Results

Predictors of non-response at wave 9 (age 55)

Among the 6,517 NCDS cohort members considered in the present analysis, 5,786 (88.8%) responded at wave 9 (age 55).

Of the 58 HES variables entering the variable selection approach (Supplementary Table S3), 10 were identified as important predictors of NCDS non-response at age 55 (Table 1). Non-response was positively associated with the proportion of OP appointments missed (odds ratio (OR) 1.03 (95% confidence interval (CI) 1.02, 1.03%) comparing those who missed all their appointments vs. those who missed none) and the number of A&E attendances (1.03 (1.01, 1.05) per A&E attendance). Almost all the selected treatments, diagnoses and operations were also positively associated with non-response, with the strongest association being with treatment for adult mental illness (1.73 (1.17, 2.51) for those ever under treatment compared to those never under treatment). The only exception was operation code H (lower digestive tract), where ever having undergone a relevant procedure was found to be protective against non-response (OR 0.73 (95% CI 0.56, 0.93)).

Restoring sample representativeness

There was strong evidence of associations between virtually all the identified HES predictors of non-response and ‘cognitive ability at age 7’ and ‘no educational qualifications at age 55’ (Supplementary Table S4). Evidence of associations with 'father in professional social class at birth’ was a little more mixed, while for ‘single and never married at age 55’ there was only evidence of association for two of the HES variables. Whilst these results suggest that this set of HES variables may not all be useful auxiliary variables for all the analysis variables, we retained them in the subsequent analyses for completeness. There was strong evidence of association between almost all the identified survey predictors of non-response and the analysis variables (Supplementary Table S5), suggesting that these represent useful auxiliary variables for the intended analyses.

Analysis A: HES linkage consenters

Cognitive ability at age 7 had mean 0.00 (95% CI -0.01, 0.01) across all NCDS cohort members (Fig. 2), with a similar distribution when restricting to the wave 9 target population (alive and still living in Great Britain). When restricting to HES linkage consenters who were eligible for linkage (lived in England for at least one wave been wave 6 and wave 9) the estimate increased to 0.12 (95% CI 0.10, 0.14), demonstrating considerable bias due to selection into linkage consent (and eligibility). Further restricting to wave 9 respondents increased the estimate to 0.16 (95% CI 0.14, 0.18), illustrating substantial bias due to non-response. Multiple imputation including only the survey predictors of non-response as auxiliary variables reduced the estimate to a level similar to that among all HES linkage consenters who were eligible for linkage (0.12; 95% CI 0.10, 0.14). Additionally including the HES predictors of non-response did not appreciably improve the estimate. Using only the

HES predictors of non-response as auxiliary variables had limited impact on restoring sample representativeness (0.15; 95% CI 0.12, 0.18).

Similar findings were observed for 'father in professional social class at birth' (Supplementary Results S1).

Analysis B: All NCDS cohort members

Cognitive ability at age 7 had mean 0.00 (95% CI -0.01, 0.01) using all available data (Fig. 3), with a similar distribution when restricting to the wave 9 target population. When additionally restricting to wave 9 respondents, there was substantial bias (0.14; 95% CI 0.13, 0.16). The MI approach using either only survey predictors of non-response (0.01; 95% CI -0.01, 0.02) or both survey and HES predictors of non-response (0.01; 95% CI -0.01, 0.03) successfully overcome this bias and restored sample representativeness.

The percentage of NCDS-comparable individuals in the population without educational qualifications was estimated to be 12.3% (95% CI 10.9%, 13.8%) using APS data (Fig. S4). Using NCDS wave 9 respondents, this was instead estimated to be 8.4% (95% CI 7.9%, 9.0%), demonstrating considerable bias relative to the population benchmark. MI estimates using survey predictors of non-response (13.7%; 95% CI 12.8%, 14.6%) or survey and HES predictors of non-response (13.7%; 95% CI 12.7%, 14.6%) were much closer to the population estimates (and with point estimates inside the population 95% CI).

Similar findings were observed for 'father in professional social class at birth' and 'single and never married at age 55' (Supplementary Results S1).

Conclusions

Summary of findings

Our analysis identified 10 HES variables associated with NCDS age 55 non-response. Most of the identified variables signified poor health, either through A&E attendances or through diagnosis of or treatment for a specific disease or condition. Whilst the existing literature on predictors of non-response in longitudinal surveys has not generally examined this area in such detail, our observations are consistent with previous findings that worse physical (2, 21-23) and mental (23-26) health are associated with subsequent non-response. There is also potential overlap with previously identified survey predictors of NCDS age 55 non-response such as self-rated general health in mid-life and conduct problems in adolescence (12).

There was generally strong evidence that the identified HES predictors of non-response were associated with the variables considered in the analyses looking to restore sample representativeness, suggesting that they may constitute useful auxiliary variables. Whilst the inclusion of HES predictors of non-response as auxiliary variables did aid in restoring sample representative to a limited extent, in analyses where the previously identified survey predictors of non-response were used there was generally no benefit of additionally including the HES variables. These results are suggestive that, for these specific variables at least, the survey predictors of non-response were sufficient to fulfil the MAR assumption, with the HES variables largely superfluous in this regard.

Strengths and limitations

There are several strengths to this analysis. This study used a large, long-running, population-representative cohort study. We utilised a data-driven approach, eschewing a theory-based approach to allow us to identify variables which may aid in maximising the plausibility of the MAR assumption without preconceptions. We explored the performance of our proposed approach to the handling of missing data through comparison to population benchmarks.

However, there are also a number of limitations. Our restriction to higher level derived diagnosis and treatment HES variables may mean that more granular relevant information was overlooked. This work concerns those born in 1958 in Great Britain – whilst it is likely to be somewhat more broadly applicable, we cannot suggest to what extent it will be generalisable to other populations. In particular, our finding that there was essentially no additional gain in sample representativeness when using HES predictors of non-response relative to analyses using only previously identified survey predictors of non-response may be specific to NCDS due to the very rich set of socio-economic and health variables available in this study. Linked administrative data may also provide useful auxiliary variables on the basis of their association with the underlying values of variable(s) subject to missingness, for example by acting as a proxy for a partially observed outcome variable (27, 28). This needs to be addressed on an analysis-specific basis so has not been considered here but is an important area for future work.

More broadly, a potential limitation of using linked administrative data in the handling of missing data lies in the nature of the linkage consent mechanism. In NCDS, opt-in linkage consent was sought at wave 8 (age 50), meaning that any cohort members who did not respond at this wave (including anyone who attritted prior to this point) will not have linked HES data – yet these individuals will constitute a large proportion of non-respondents at subsequent waves, so are a subgroup for whom appropriate non-response handling is essential. This emphasises the importance of early (ideally at study initiation) opt-in linkage consents or alternative (e.g. opt-out) consent mechanisms to allow access to linked data for as many study participants as possible. It also highlights the potential of surveys which utilise an administrative data sampling frame, meaning that some administrative data should be available for all sampled individuals, including baseline non-respondents, allowing particularly thorough investigation of non-response.

Implications for analyses using NCDS data

We have demonstrated that principled methods for missing data handling (in this case MI) utilising appropriately chosen auxiliary variables have the ability to restore sample representativeness in NCDS. Whilst the inclusion of HES predictors of non-response did aid in restoring sample representative to a limited extent, previously identified survey predictors of non-response were far more important. For users of NCDS data, we therefore emphasise previous guidance on the inclusion of appropriately chosen survey predictors of non-response in analyses (12, 29) and do not suggest the default inclusion of HES variables on the basis of their association with non-response. However, as noted, auxiliary variables should also be considered based on their association with the underlying values of variable(s) subject to missingness, and HES variables may therefore be relevant for certain analyses.

Conclusions

In this analysis we explored the extent to which administrative (HES) data could aid in predicting survey (NCDS) non-response and restoring survey sample representativeness. Whilst the inclusion of HES predictors of non-response did aid in restoring sample representative to a limited extent, previously identified survey predictors of non-response were the far more important, highlighting their value in analyses of data subject to missingness.

Ethics approval

NCDS was approved by the National Health Service Research Ethics Committee and all participants have given informed consent.

Data availability

NCDS data (SN 2000032), including linked NCDS-HES data (SN 8697), are available through the UK Data Service.

Funding

This work was supported by the Economic & Social Research Council and Administrative Data Research UK [grant number ES/V006037/1] and the Centre for Longitudinal Studies is supported by the Economic & Social Research Council [grant number ES/W013142/1]. 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. This work was in part supported by the NIHR through the Great Ormond Street Hospital Biomedical Research Centre.

REFERENCES

1. Watson N, Wooden M. Identifying factors affecting longitudinal survey response. In: Lynn P, editor. Methodology of Longitudinal Surveys. Chichester: Wiley; 2009. p. 157-82.
2. Galea S, Tracy M. Participation Rates in Epidemiologic Studies. Annals of Epidemiology. 2007;17(9):643-53.
3. Carpenter JR, Kenward MG. Multiple Imputation and its Application. Chichester, UK: John Wiley & Sons, Ltd; 2013.
4. 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.
5. Seaman SR, White IR. Review of inverse probability weighting for dealing with missing data. Stat Methods Med Res. 2013;22(3):278-95.
6. Little RJA, Rubin DB. Statistical Analysis with Missing Data. Second Edition. Hoboken, NJ: Wiley; 2002.
7. Calderwood L, Lessof C. Enhancing Longitudinal Surveys by Linking to Administrative Data. In: Lynn P, editor. Methodology of Longitudinal Surveys. Chichester: Wiley; 2009. p. 55-72.
8. Power C, Elliott J. Cohort profile: 1958 British birth cohort (National Child Development Study). Int J Epidemiol. 2006;35(1):34-41.
9. NHS Digital. Hospital Episode Statistics (HES). 2020 [Available from: https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/hospital-episode-statistics. Accessed 19 July 2022.].
10. Kerry-Barnard S, Gomes D. National Child Development Study: A guide to the linked health administrative datasets – Hospital Episode Statistics (HES). London: UCL Centre for Longitudinal Studies; 2020.
11. University College London, UCL Institute of Education, Centre for Longitudinal Studies, NHS Digital. National Child Development Study: Linked Health Administrative Datasets (Hospital Episode Statistics), England, 1997-2017: Secure Access. [data collection]. UK Data Service. SN: 8697, DOI: 10.5255/UKDA-SN-8697-1. 2021.
12. Mostafa T, Narayanan M, Pongiglione B, Dodgeon B, Goodman A, Silverwood RJ, et al. Missing at random assumption made more plausible: evidence from the 1958 British birth cohort. J Clin Epidemiol. 2021;136:44-54.
13. Archer G, Xun WW, Stuchbury R, Nicholas O, Shelton N. Are 'healthy cohorts' real-world relevant? Comparing the National Child Development Study (NCDS) with the ONS Longitudinal Study (LS). Longitudinal and Life Course Studies. 2020;11(3):307–30.
14. Silverwood RJ, Goodman A, Ploubidis GB. Letter to the editor: Don't forget survey data: 'healthy cohorts' are 'real-world' relevant if missing data are handled appropriately. Longitudinal and Life Course Studies. 2022;13(2):335-41.
15. Herbert A, Wijlaars L, Zylbersztejn A, Cromwell D, Hardelid P. Data Resource Profile: Hospital Episode Statistics Admitted Patient Care (HES APC). International Journal of Epidemiology. 2017;46(4):1093-i.
16. Silverwood RJ, Rajah N, Calderwood L, De Stavola BL, Harron K, Ploubidis GB. Examining the linkage quality and sample representativeness of linked National Child Development Study and Hospital Episode Statistics data. CLS Working Paper 2022/5. London: UCL Centre for Longitudinal Studies; 2022.
17. Office for National Statistics. Annual population survey (APS) QMI 2022 [Available from:

https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/employmentandemployeetypes/methodologies/annualpopulationsurveyapsqmi. Accessed 19 July 2022].
18. Office for National Statistics, Social Survey Division. Annual Population Survey, January - December, 2013. [data collection]. 11th Edition. UK Data Service. SN: 7536, DOI: 10.5255/UKDA-SN-7536-11.2020.
19. Tibshirani R. Regression Shrinkage and Selection Via the Lasso. Journal of the Royal Statistical Society: Series B (Methodological). 1996;58(1):267-88.
20. White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. Stat Med. 2011;30(4):377-99.
21. Young AF, Powers JR, Bell SL. Attrition in longitudinal studies: who do you lose? Australian and New Zealand Journal of Public Health. 2006;30(4):353-61.
22. Banks J, Muriel A, Smith JP. Attrition and health in ageing studies: Evidence from ELSA and HRS. Longit Life Course Stud. 2011;2(2).
23. Tyrrell J, Zheng J, Beaumont R, Hinton K, Richardson TG, Wood AR, et al. Genetic predictors of participation in optional components of UK Biobank. Nature Communications. 2021;12(1):886.
24. Cornish RP, Macleod J, Boyd A, Tilling K. Factors associated with participation over time in the Avon Longitudinal Study of Parents and Children: a study using linked education and primary care data. International Journal of Epidemiology. 2021;50(1):293-302.
25. Taylor AE, Jones HJ, Sallis H, Euesden J, Stergiakouli E, 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.
26. Fröjd SA, Kaltiala-Heino R, Marttunen MJ. Does problem behaviour affect attrition from a cohort study on adolescent mental health? European Journal of Public Health. 2011;21(3):306-10.
27. Cornish RP, Tilling K, Boyd A, Davies A, Macleod J. Using linked educational attainment data to reduce bias due to missing outcome data in estimates of the association between the duration of breastfeeding and IQ at 15 years. International Journal of Epidemiology. 2015;44(3):937-45.
28. Cornish RP, Macleod J, Carpenter JR, Tilling K. Multiple imputation using linked proxy outcome data resulted in important bias reduction and efficiency gains: a simulation study. Emerging Themes in Epidemiology. 2017;14(1):14.
29. Silverwood R, Narayanan M, Dodgeon B, Ploubidis G. Handling missing data in the National Child Development Study: User Guide (Version 2). London: UCL Centre for Longitudinal Studies; 2021.

TABLES AND FIGURES

Table 1. Estimated odds ratios and 95% confidence intervals for identified Hospital Episode Statistics (HES) predictors of non-response at sweep 9 (age 55) in the 1958 British National Child Development Study.

HES variable[A]	Odds ratio	95% confidence interval
Number of A&E attendances (per unit increase)	1.03	1.01, 1.05
Proportion of OP appointments missed (per unit increase)	1.03	1.02, 1.03
Treatment by Adult Mental Illness	1.73	1.17, 2.51
ICD Chapter IV: Endocrine, nutritional and metabolic diseases	1.17	0.91, 1.51
ICD Chapter V: Mental and behavioural disorders	1.13	0.83, 1.53
ICD Chapter VI: Diseases of the nervous system	1.14	0.85, 1.51
ICD Chapter X: Diseases of the respiratory system	1.20	0.93, 1.53
ICD Chapter XVIII: Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified	1.20	0.98, 1.45
Operation code H: Lower digestive tract	0.73	0.56, 0.93
Operation code T: Soft tissue	1.21	0.95, 1.53

[A] Unless otherwise noted in column 'HES variables', the reference category is having not been diagnosed or treated for the relevant condition.
A&E: accident and emergency; OP: outpatients.

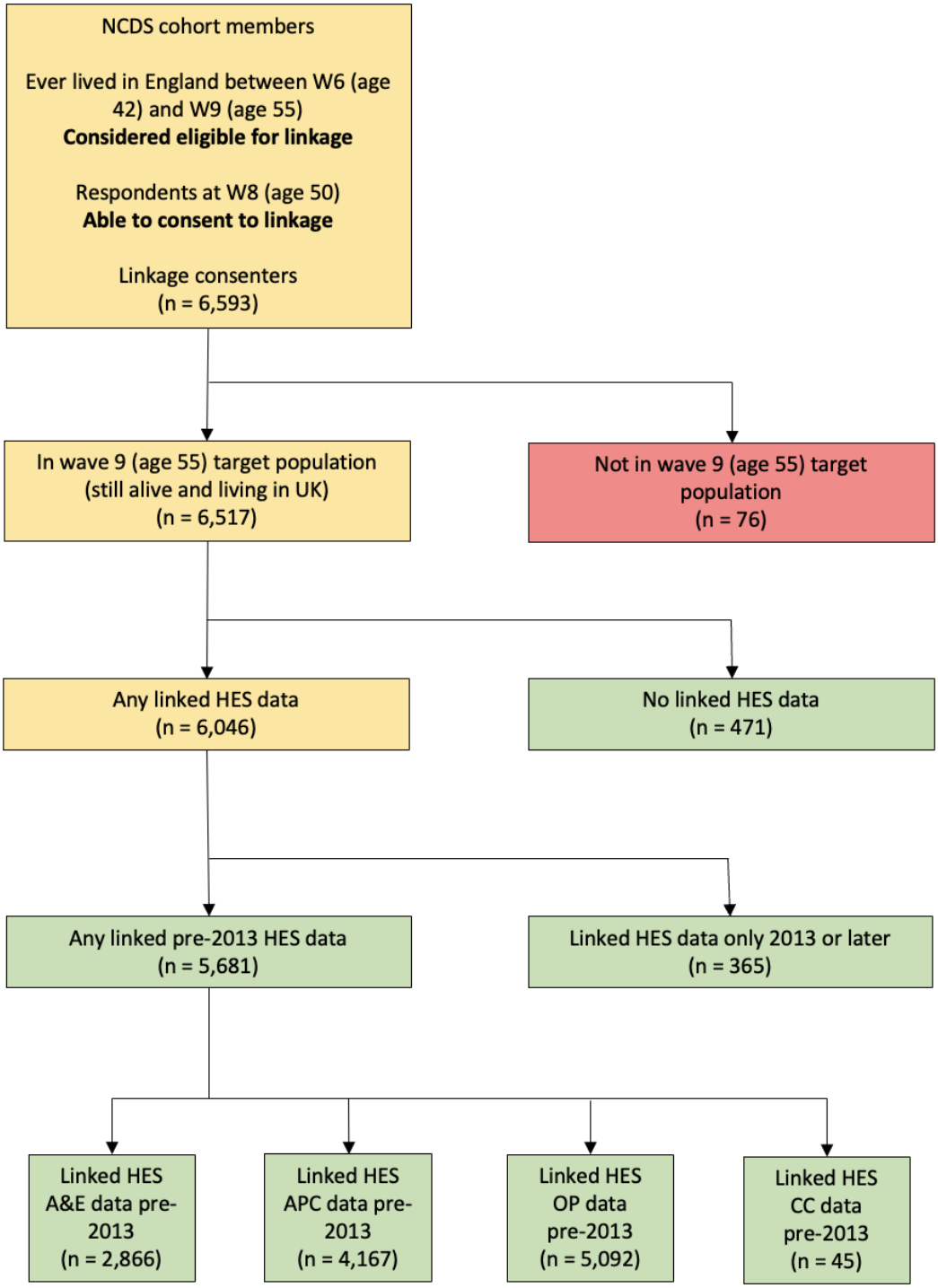

Figure 1. Flow diagram showing 1958 British National Child Development Study-Hospital Episode Statistics data linkage and data availability. APC: admitted patient care; CC: critical care; A&E: accident and emergency; OP: outpatients.

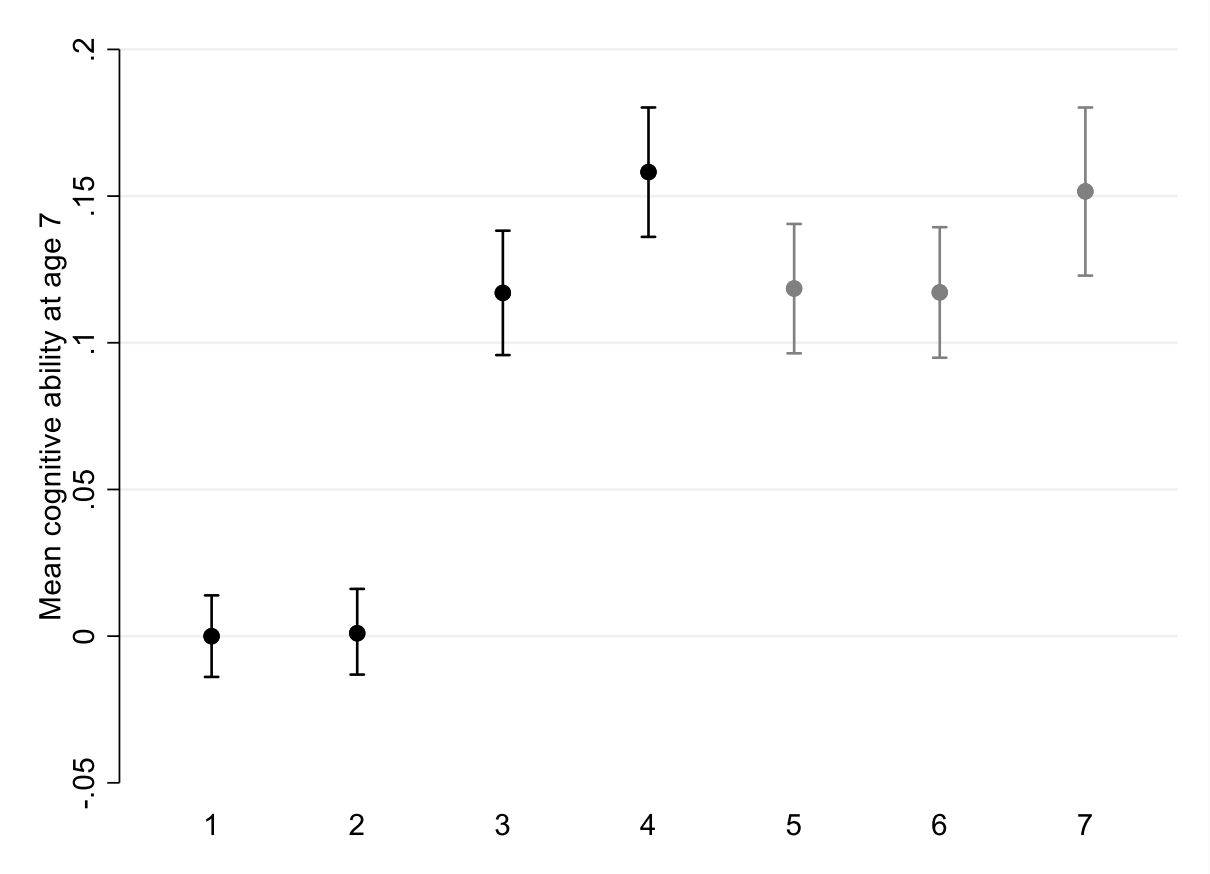

Figure 2. Mean (95% confidence interval) cognitive ability at age 7 in the National Child Development Study before and after handling missing data (Analysis A).
Analysis 1: Distribution using all available data (n = 14,407).
Analysis 2: Distribution restricted to wave 9 target population (alive and still living in GB) (n = 12,938).
Analysis 3: Distribution restricted to wave 9 target population and HES linkage consenters who were eligible for linkage (lived in England for at least one wave been W6 and W9) (n = 5,546).
Analysis 4: Distribution restricted to wave 9 respondents within HES linkage consenters who were eligible for linkage (n = 4,928).
Analysis 5: MI analysis using only selected survey predictors of non-response as auxiliary variables, restricted to wave 9 target population, HES linkage consenters who were eligible for linkage and those non-missing for the variable of interest, using information on the variable of interest from wave 9 respondents only (n = 5,546).
Analysis 6: MI analysis using selected HES predictors of non-response in addition to selected survey predictors of non-response as auxiliary variables, restricted to wave 9 target population, HES linkage consenters who were eligible for linkage and those non-missing for the variable of interest, using information on the variable of interest from wave 9 respondents only (n = 5,546).
Analysis 7: MI analysis using only selected HES predictors of non-response as auxiliary variables, restricted to wave 9 target population, HES linkage consenters who were eligible for linkage and those non-missing for the variable of interest, using information on the variable of interest from wave 9 respondents only (n = 5,546).

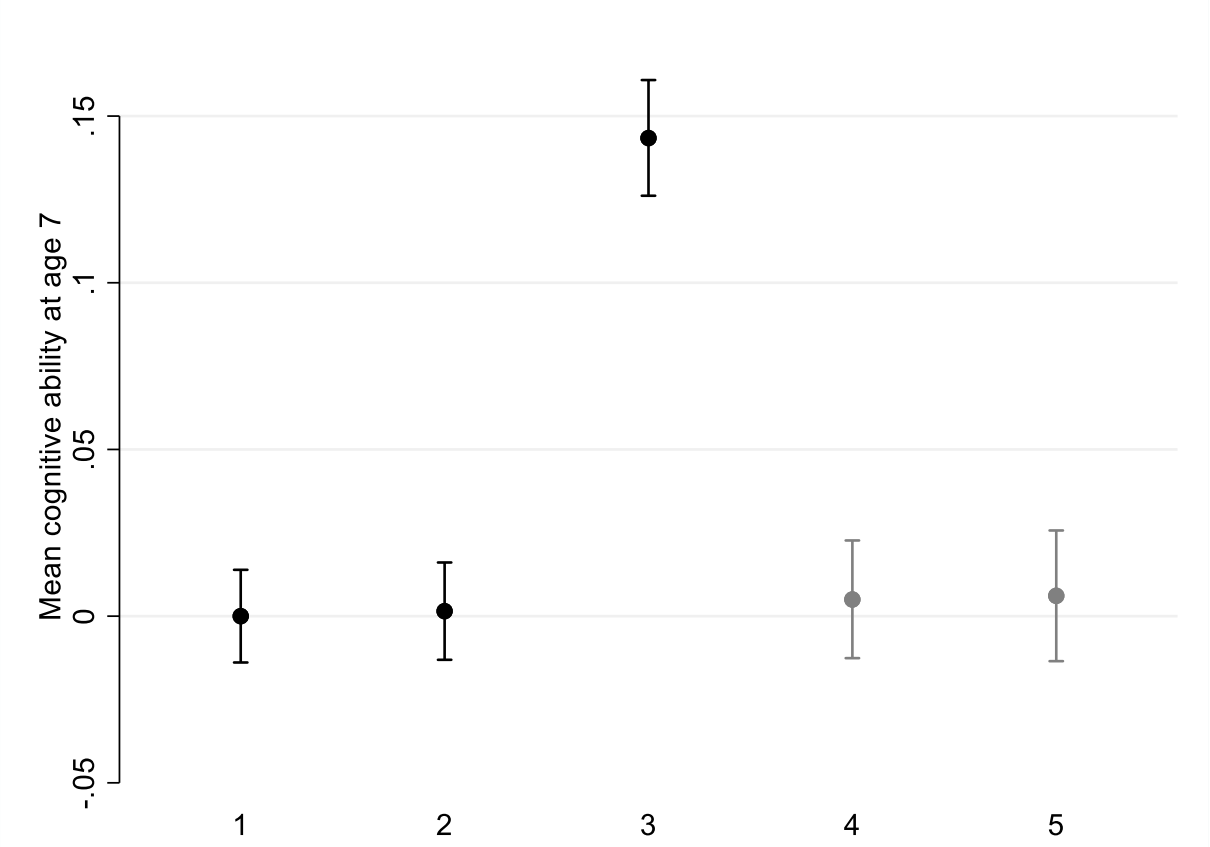

Figure 3. Mean (95% confidence interval) cognitive ability at age 7 in the National Child Development Study before and after handling missing data (Analysis B).
Analysis 1: Distribution using all available data (n = 14,407).
Analysis 2: Distribution restricted to wave 9 target population (alive and still living in GB) (n = 12,938).
Analysis 3: Distribution restricted to wave 9 respondents (n = 7,839).
Analysis 4: MI analysis using only selected survey predictors of non-response as auxiliary variables, restricted to wave 9 target population and those non-missing for the variable of interest, using information on the variable of interest from wave 9 respondents only (n = 12,938).
Analysis 5: MI analysis using selected HES predictors of non-response in addition to selected survey predictors of non-response as auxiliary variables, restricted to wave 9 target population and those non-missing for the variable of interest, using information on the variable of interest from wave 9 respondents only (n = 12,938).

SUPPLEMENTARY MATERIAL

Contents

Methods S1. Secondary variable selection analysis.

Methods

In a secondary analysis we used a multi-stage p-value-based variable selection approach, similar to that employed by Mostafa et al (1), for comparison with the primary approach using the LASSO. We employed a modified Poisson model with robust standard errors (2) to estimate risk ratios at each stage:

> Stage 1: Univariable modified Poisson regressions of non-response at age 55 on each identified predictor from the HES datasets. Retain variables where $P < 0.05$.
>
> Stage 2: Multivariable modified Poisson regression of non-response at age 55 on all predictors retained from stage 1. Retain variables where $P < 0.05$

In the interests of transparency, we note that this approach was initially planned to be our primary analysis with the LASSO a secondary analysis, but we have switched their reporting due to poor performance (in terms of number of predictors of non-response identified) of the p-value-based approach. As the ultimate aim of the present analysis was to identify a set of variables containing sufficient information with respect to NCDS age 55 non-response to act as useful auxiliary variables in subsequent analyses, we considered this change in methodological focus to be necessary and do not believe that it adversely affects the interpretation of our findings.

Results

Fifty-eight variables were derived from HES data and formed the input to stage 1 of the process to identify predictors of non-response at age 55. In the secondary analysis, 22 of these were found to have univariable associations with non-response, so passed from stage 1 to stage 2. In stage 2, three of these variables remained associated with non-response in the multivariable model so were identified as important HES predictors of non-response:

1. Proportion of outpatient appointments missed
2. Number of A&E attendances
3. Treatment for adult mental illness (ever treated)

Since only a small number of predictors of non-response was identified using this approach, and since they were a subset of those identified in the primary analysis using LASSO, these variables were not considered further in isolation.

References

1. Mostafa T, Narayanan M, Pongiglione B, Dodgeon B, Goodman A, Silverwood RJ, et al. Missing at random assumption made more plausible: evidence from the 1958 British birth cohort. J Clin Epidemiol. 2021;136:44-54.

2. Zou G. A modified poisson regression approach to prospective studies with binary data. Am J Epidemiol. 2004;159(7):702-6.

Methods S2. Sample representativeness analyses.

We undertook a number of different analyses to assess how effective the identified HES predictors of NCDS age 55 non-response were at restoring sample representativeness despite selective attrition. We considered sample representativeness in terms of two NCDS variables observed in early life (father's social class at birth and cognitive ability at age 7) and in terms of two NCDS variables in observed in later life (educational qualifications at wave 9 (age 55) and marital status at wave 9 (age 55)). We undertook two different analyses to explore different aspects of restoring sample representativeness, the first ("Analysis A") considering only HES linkage consenters and the second ("Analysis B") considering all NCDS cohort members. All four NCDS variables were considered within Analysis B, but only the two early life variables within Analysis A. Full details of the analyses are provided below.

Analysis A: HES linkage consenters

This analysis focuses on HES linkage consenters who were eligible for linkage (lived in England for at least one wave been wave 6 and wave 9) and who were within the wave 9 target population. These individuals are non-missing for all HES variables, therefore it is possible to conduct MI analyses where the only auxiliary variables are the selected HES predictors of non-response, allowing an assessment of whether the HES variables in isolation are helpful in restoring sample representativeness. However, analyses restricted to HES linkage consenters form only a limited proportion of analyses that could be undertaken using NCDS data and results in this setting may not be more broadly applicable. As these analyses relate to a subpopulation which would be impossible to identify within the APS (individuals who would hypothetically consent to HES linkage), these analyses were restricted to restoring sample representativeness of variables observed in early life.

Restoring sample representativeness of NCDS variables in observed in early life (father's social class at birth and cognitive ability at age 7): wave 9 respondents vs. complete distribution from early life

Analyses:

1. Distribution using all available data
2. Distribution restricted to wave 9 target population (alive and still living in GB)
3. Distribution restricted to wave 9 target population and HES linkage consenters who were eligible for linkage (lived in England for at least one wave been W6 and W9)
4. Distribution restricted to wave 9 respondents within HES linkage consenters who were eligible for linkage
5. MI analysis using only selected survey predictors of non-response as auxiliary variables, restricted to wave 9 target population, HES linkage consenters who were eligible for linkage and those non-missing for the variable of interest, using information on the variable of interest from wave 9 respondents only
6. MI analysis using selected HES predictors of non-response in addition to selected survey predictors of non-response as auxiliary variables, restricted to wave 9 target population, HES linkage consenters who were eligible for linkage and those non-

missing for the variable of interest, using information on the variable of interest from wave 9 respondents only

7. MI analysis using only selected HES predictors of non-response as auxiliary variables, restricted to wave 9 target population, HES linkage consenters who were eligible for linkage and those non-missing for the variable of interest, using information on the variable of interest from wave 9 respondents only

Comparisons of interest:

- 2 vs. 1 shows any difference between wave 9 target population and all available data
- 3 vs 2 shows any bias introduced by only considering HES linkage consenters who were eligible for linkage (among the wave 9 target population)
- 4 vs. 3 shows any bias due to non-response at wave 9 (among the wave 9 target population and HES linkage consenters who were eligible for linkage)
- 5 vs. 4 shows to what extent sample representativeness can be restored using only the selected survey predictors of non-response (among the wave 9 target population and HES linkage consenters who were eligible for linkage)
- 6 vs. 4 shows to what extent sample representativeness can be restored using both the selected HES predictors of non-response and the selected survey predictors of non-response (among the wave 9 target population and HES linkage consenters who were eligible for linkage)
- 6 vs. 5 shows the added value of the selected HES predictors of non-response relative to only the selected survey predictors of non-response for restoring sample representativeness (among the wave 9 target population and HES linkage consenters who were eligible for linkage)
- 7 vs. 4 shows to what extent sample representativeness can be restored using only the selected HES predictors of non-response (among the wave 9 target population and HES linkage consenters who were eligible for linkage

Analysis B: All NCDS cohort members

This analysis focuses on all NCDS cohort members (within the wave 9 target population). This includes individuals who did not consent to HES linkage (or who did consent but were ineligible for linkage) and are therefore missing for all HES variables, making it impossible to conduct MI analyses where the only auxiliary variables are the selected HES predictors of non-response, meaning an assessment of whether the HES variables in isolation are helpful in restoring sample representativeness is not possible. However, analyses of all NCDS cohort members will be more commonly undertaken than those restricted to HES linkage consenters who were eligible for linkage (Analysis A) so there is much interest in restoring sample representativeness in this setting. As these analyses relate to the whole NCDS sample (which should be representative of the population), corresponding population statistics APS can be obtained from APS and analyses related to restoring sample representativeness of both early life and later life NCDS variables are undertaken, though these were by necessity structured slightly differently.

Restoring sample representativeness of NCDS variables in observed in early life (father's social class at birth and cognitive ability at age 7): wave 9 respondents vs. complete distribution from early life

Analyses:

1. Distribution using all available data
2. Distribution restricted to wave 9 target population (alive and still living in GB)
3. Distribution restricted to wave 9 respondents
4. MI analysis using only selected survey predictors of non-response as auxiliary variables, restricted to wave 9 target population and those non-missing for the variable of interest, using information on the variable of interest from wave 9 respondents only
5. MI analysis using selected HES predictors of non-response in addition to selected survey predictors of non-response as auxiliary variables, restricted to wave 9 target population and those non-missing for the variable of interest, using information on the variable of interest from wave 9 respondents only

Comparisons of interest:

- 2 vs. 1 shows any difference between wave 9 target population and all available data
- 3 vs. 2 shows any bias due to non-response at wave 9 (among the wave 9 target population)
- 4 vs. 3 shows to what extent sample representativeness can be restored using only the selected survey predictors of non-response (among the wave 9 target population)
- 5 vs. 3 shows to what extent sample representativeness can be restored using both the selected HES variables and the selected survey predictors of non-response (among the wave 9 target population)
- 5 vs. 4 shows the added value of the selected HES predictors of non-response relative to only the selected survey predictors of non-response for restoring sample representativeness (among the wave 9 target population)

Restoring sample representativeness of NCDS variables in observed in later life (marital status at wave 9 (age 55) and educational qualifications at wave 9 (age 55)): wave 9 respondents vs. population benchmark data from APS

Analyses:

1. Distribution using population benchmark data from APS
2. Distribution among wave 9 respondents (which in this case is all available information)
3. MI analysis using only selected survey predictors of non-response as auxiliary variables, restricted to wave 9 target population
4. MI analysis using selected HES predictors of non-response in addition to selected survey predictors of non-response as auxiliary variables, restricted to wave 9 target population

Comparisons of interest:

- 2 vs. 1 shows any bias introduced by selective response at wave 9 respondents (among the wave 9 target population)
- 3 vs. 2 shows to what extent sample representativeness can be restored using only the selected survey predictors of non-response
- 4 vs. 2 shows to what extent sample representativeness can be restored using both the selected HES variables and the selected survey predictors of non-response
- 4 vs. 3 shows the added value of the selected HES predictors of non-response relative to only the selected survey predictors of non-response for restoring sample representativeness

Results S1. Sample representativeness analyses.

Analysis A: HES linkage consenters

Using all available data, 4.5% of cohort members had a father in the professional social class (95% CI 4.2%, 4.9%) (Fig. S2), with a similar estimate within the wave 9 target population. Amongst the HES linkage consenters who were eligible for linkage the estimate increased slightly to 5.0% (95% CI 4.5%, 5.6%), demonstrating moderate bias due to selection into linkage consent. When considering only wave 9 respondents the estimate was 5.3% (95% CI 4.7%, 6.0%), showing limited bias due to non-response. None of the MI approaches overcame this limited non-response bias, with similar estimates when using only survey predictors of non-response (5.4%; 95% CI 4.8%, 6.0%), both survey and HES predictors of non-response (5.4%; 95% CI 4.8%, 6.0%), or only HES predictors of non-response (5.3%; 95% CI 4.7%, 6.0%).

Analysis B: All NCDS cohort members

The proportion of cohort members whose father was in the professional social class was 4.5% (95% CI 4.2%, 4.9%) using all available data (Fig. S3), with a similar distribution when restricting to the wave 9 target population (4.6%; 95% CI 4.2%, 4.9%). There was evidence of bias (non-overlapping CIs) when restricting to wave 9 respondents (5.4%; 95% CI 5.0%, 6.0%), though this was overcome using MI utilising either the survey predictors of non-response alone (4.7%; 95% CI 4.3%, 5.2%) or in combination with the HES predictors of non-response (4.7%; 95% CI 4.3%, 5.3%).

The APS-derived estimated prevalence of being single and never been married of 11.4% had a relatively wide 95% CI (10.0%, 12.8%) (Fig. S5). The corresponding estimate among NCDS wave 9 respondents 10.0% (95% CI 9.3%, 10.6%), showing relative bias in the point estimate even if there remained overlap in the 95% CIs. The MI analysis using only survey predictors of non-response overcame most of this bias (10.9%; 95% CI 10.2%, 11.7%), though the combination of survey and HES predictors of non-response performed even better (11.5%; 95% CI 10.8%, 12.4%).

Table S1. Survey-based predictors of National Child Development Study non-response at wave 9 (age 55).

Sweep	Variable
Sweep 0 (age 0)	Mother's age
	Number of persons per room
	Parity
	Social class of mother's father when she left school
	Sex of child
	Social class of mother's husband
Sweep 1 (age 7)	Dad stayed on at school after minimum age
	Social problems (alcoholism etc.)
	Cognitive ability summary
	Ever breastfed
Sweep 2 (age 11)	Cognitive ability summary
	Conduct problems
Sweep 3 (age 16)	Child receiving help at school – backwardness
	Child's school attendance
	How long since child drank alcohol
	Test 2 – mathematics comprehension
	Conduct problems
Sweep 4 (age 23)	Legal marital status
	Voted in 1979 general election
Sweep 5 (age 33)	Telephone in home
	How much physical effort in job
	Voted in 1987 general election
	Housing tenure
	Social capital score (people turn to for advice, support)
Sweep 6 (age 42)	Membership in organisations
Biomedical sweep (age 44)	Self-rated general health
Sweep 7 (age 46)	Marital status - de facto
Sweep 8 (age 50)	Total number of natural children
	Employer provided pension scheme
	Non-response at sweeps 1-8

Table S2. Derivation of Hospital Episode Statistics variables for consideration as potential predictors of National Child Development Study non-response.

Variable	Number of variables	Details	HES dataset	Period
Number of A&E attendances	1	Count of the number of times the individual has had an A&E appointment.	A&E	2007-2013
Proportion of A&E investigations to attendances	1	The individual's average number of investigations in A&E divided by the number of A&E attendances the individual has.	A&E	2007-2013
Average number of A&E treatments	1	The average number of treatments the individual received in A&E, which is the number of treatments divided by the number of A&E visits.	A&E	2007-2013
Number of APC spells	1	Count of the number of spells an individual has, defined by number of unique admission dates.	APC	1997-2013
Number of OP appointments	1	Count of the number of appointments in the outpatient data.	OP	2003-2013
Percentage of OP appointments missed	1	The number of outpatient appointments missed divided by the total number of outpatient appointments the individual has. Only derived for individuals with a minimum of two outpatient appointments.	OP	2003-2013
OP treatment (High Dependency Care, Intensive Care, Oncologist, Rehabilitation, Adult Mental Illness, Cardiology, Plastic Surgery)	7	Whether the individual had ever been treated under the relevant area at any point; 0 if they had not and 1 if they had.	OP	2003-2013
Diagnosis codes (A-Z)	21	Whether the individual had the diagnosis code (ICD-10) recorded at any point: 0 if they had not, 1 if they had.	APC	1997-2013
Operation codes (A-Z)	24	Whether the individual had the operation code (OPCS-4) recorded at any point: 0 if they had not, 1 if they had.	APC	1997-2013
Total	**58**			

HES: Hospital Episode Statistics; APC: admitted patient care; CC: critical care; A&E: accident and emergency; OP: outpatients; ICD-10: International Statistical Classification of Diseases and Related Health Problems 10th Revision; OPCS-4: Office of Population Censuses and Surveys Classification of Interventions and Procedures version 4.

Note: We assume that cohort members who were eligible for and consented to linkage but did not have linked data for a given HES dataset truly did not have a relevant interaction (e.g., admission, outpatient appointment) with an NHS hospital in England over the corresponding time period. Such individuals are therefore included in the analysis with HES variables derived to reflect that, for example, if they had no linked HES APC data then they did not receive any APC-based diagnoses or undergo any APC-based treatments.

Table S3. Descriptive statistics for Hospital Episode Statistics (HES) variables used in the analysis.

	Overall (n = 6517)	Respondents (n = 5786)	Non-respondents (n = 731)
HES variable	Mean (SD)	Mean (SD)	Mean (SD)
Number of A&E attendances	1.5 (3.1)	1.4 (2.9)	2.2 (4.1)
Proportion of A&E investigations to attendances	0.3 (0.9)	0.3 (0.9)	0.3 (0.8)
Average number of A&E treatments	0.6 (1.2)	0.6 (1.2)	0.7 (1.1)
Number of APC spells	2.4 (4.4)	2.3 (4.1)	3.1 (6.0)
Number of OP appointments	11.1 (18.6)	10.8 (17.7)	13.9 (24.4)
Percentage of OP appointments missed	3.9 (9.2)	3.5 (8.5)	7.2 (13.0)
HES variable	n (% of total)	n (% of those in HES variable stratum)	n (% of those in HES variable stratum)
Treatment by High Dependency Care			
No	6466 (99.2%)	5742 (88.8%)	724 (11.2%)
Yes	51 (0.8%)	44 (86.3%)	7 (13.7%)
Treatment by Intensive Care			
No	6498 (99.7%)	5769 (88.8%)	729 (11.2%)
Yes	19 (0.3%)	17 (89.5%)	2 (10.5%)
Treatment by Oncologist			
No	6440 (98.8%)	5719 (88.8%)	721 (11.2%)
Yes	77 (1.2%)	67 (87.0%)	10 (13.0%)
Treatment by Rehabilitation			
No	6450 (99.0%)	5729 (88.8%)	721 (11.2%)
Yes	67 (1.0%)	57 (85.1%)	10 (14.9%)
Treatment by Adult Mental Illness			
No	6337 (97.2%)	5652 (89.2%)	685 (10.8%)
Yes	180 (2.8%)	134 (74.4%)	46 (25.6%)
Treatment by Cardiology			
No	5829 (99.4%)	5196 (89.1%)	633 (10.9%)
Yes	688 (10.6%)	590 (85.8%)	98 (14.2%)
Treatment by Plastic Surgery			
No	6259 (96.0%)	5561 (88.8%)	698 (11.2%)
Yes	258 (4.0%)	225 (87.2%)	33 (12.8%)
ICD Chapter I: Certain infectious and parasitic diseases			
No	6254 (96.0%)	5566 (89.0%)	688 (11.0%)
Yes	263 (4.0%)	220 (83.7%)	43 (16.3%)
ICD Chapter II: Neoplasms			

No	5940 (91.1%)	5279 (88.9%)	661 (11.1%)
Yes	577 (8.9%)	507 (87.9%)	70 (12.1%)
ICD Chapter III: Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism			
No	6323 (97.0%)	5618 (88.9%)	705 (11.1%)
Yes	194 (3.0%)	168 (86.6%)	26 (13.4%)
ICD Chapter IV: Endocrine, nutritional and metabolic diseases			
No	5894 (90.4%)	5266 (89.3%)	628 (10.7%)
Yes	623 (9.6%)	520 (83.5%)	103 (16.5%)
ICD Chapter V: Mental and behavioural disorders			
No	6100 (93.6%)	5448 (89.3%)	652 (10.7%)
Yes	417 (6.4%)	338 (81.1%)	79 (18.9%)
ICD Chapter VI: Diseases of the nervous system			
No	6064 (93.0%)	5406 (89.1%)	658 (10.9%)
Yes	453 (7.0%)	380 (83.9%)	73 (16.1%)
ICD Chapter VII: Diseases of the eye and adnexa			
No	6296 (96.6%)	5599 (88.9%)	697 (11.1%)
Yes	221 (3.4%)	187 (84.6%)	34 (15.4%)
ICD Chapter VIII: Diseases of the ear and mastoid process			
No	6425 (98.6%)	5709 (88.9%)	716 (11.1%)
Yes	92 (1.4%)	77 (83.7%)	15 (16.3%)
ICD Chapter IX: Diseases of the circulatory system			
No	5412 (83.0%)	4816 (89%)	596 (11%)
Yes	1105 (17.0%)	970 (87.8%)	135 (12.2%)
ICD Chapter X: Diseases of the respiratory system			
No	5865 (90.0%)	5240 (89.3%)	625 (10.7%)
Yes	652 (10.0%)	546 (83.7%)	106 (16.3%)
ICD Chapter XI: Diseases of the digestive system			
No	4956 (76.0%)	4412 (89.0%)	544 (11.0%)
Yes	1561 (24.0%)	1374 (88.0%)	187 (12.0%)
ICD Chapter XII: Diseases of the skin and subcutaneous tissue			
No	6091 (93.5%)	5404 (88.7%)	687 (11.3%)
Yes	426 (6.5%)	382 (89.7%)	44 (10.3%)

ICD Chapter XIII: Diseases of the musculoskeletal system and connective tissue			
No	5352 (82.1%)	4782 (89.3%)	570 (10.7%)
Yes	1165 (17.9%)	1004 (86.2%)	161 (13.8%)
ICD Chapter XIV: Diseases of the genitourinary system			
No	5322 (81.7%)	4739 (89%)	583 (11%)
Yes	1195 (18.3%)	1047 (87.6%)	148 (12.4%)
ICD Chapter XV: Pregnancy, childbirth and the puerperium			
No	6336 (97.2%)	5625 (88.8%)	711 (11.2%)
Yes	181 (2.8%)	161 (89%)	20 (11%)
ICD Chapter XVII: Congenital malformations, deformations and chromosomal abnormalities			
No	6504 (99.8%)	5773 (88.8%)	731 (11.2%)
Yes	13 (0.2%)	13 (100%)	0 (0%)
ICD Chapter XVIII: Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified			
No	4998 (76.7%)	4489 (89.8%)	509 (10.2%)
Yes	1519 (23.3%)	1297 (85.4%)	222 (14.6%)
ICD Chapter XIX: Injury, poisoning and certain other consequences of external causes			
No	5819 (89.3%)	5195 (89.3%)	624 (10.7%)
Yes	698 (10.7%)	591 (84.7%)	107 (15.3%)
ICD Chapter XX: External causes of morbidity and mortality			
No	6073 (93.2%)	5415 (89.2%)	658 (10.8%)
Yes	444 (6.8%)	371 (83.6%)	73 (16.4%)
ICD Chapter XXI: Factors influencing health status and contact with health services			
No	4658 (71.5%)	4154 (89.2%)	504 (10.8%)
Yes	1859 (28.5%)	1632 (87.8%)	227 (12.2%)
ICD Chapter XXII: Codes for special purposes			
No	6512 (99.9%)	5782 (88.8%)	730 (11.2%)
Yes	5 (0.1%)	4 (80%)	1 (20%)
Operation code A – Nervous system			
No	6130 (94.1%)	5449 (88.9%)	681 (11.1%)
Yes	387 (5.9%)	337 (87.1%)	50 (12.9%)
Operation code B – Endocrine system			
No	6325 (97.0%)	5616 (88.8%)	709 (11.2%)

Yes	192 (3.0%)	170 (88.5%)	22 (11.5%)
Operation code C - Eye			
No	6306 (96.8%)	5606 (88.9%)	700 (11.1%)
Yes	211 (3.2%)	180 (85.3%)	31 (14.7%)
Operation code D – Ear			
No	6492 (99.6%)	5764 (88.8%)	728 (11.2%)
Yes	25 (0.4%)	22 (88.0%)	3 (12.0%)
Operation code E – Respiratory tract			
No	6214 (95.4%)	5525 (88.9%)	689 (11.1%)
Yes	303 (4.6%)	261 (86.1%)	42 (13.9%)
Operation Code F – Mouth			
No	6216 (95.4%)	5521 (88.8%)	695 (11.2%)
Yes	301 (4.6%)	265 (88.0%)	36 (12.0%)
Operation code G – Upper digestive tract			
No	5805 (89.1%)	5161 (88.9%)	644 (11.1%)
Yes	712 (10.9%)	625 (87.8%)	87 (12.2%)
Operation code H – Lower digestive tract			
No	5684 (87.2%)	5042 (88.7%)	642 (11.3%)
Yes	833 (12.8%)	744 (89.3%)	89 (10.7%)
Operation code J – Other abdominal organs (principally digestive)			
No	6348 (97.4%)	5641 (88.9%)	707 (11.1%)
Yes	169 (2.6%)	145 (85.8%)	24 (14.2%)
Operation code K – Heart			
No	6292 (96.5%)	5597 (89.0%)	695 (11.0%)
Yes	225 (3.5%)	189 (84.0%)	36 (16.0%)
Operation code L – Arteries and veins			
No	6291 (96.5%)	5586 (88.8%)	705 (11.2%)
Yes	226 (3.5%)	200 (88.5%)	26 (11.5%)
Operation code M – Urinary			
No	6075 (93.2%)	5408 (89%)	667 (11%)
Yes	442 (6.8%)	378 (85.5%)	64 (14.5%)
Operation code N – Male genital organs			
No	6318 (96.9%)	5603 (88.7%)	715 (11.3%)
Yes	199 (3.1%)	183 (92%)	16 (8%)
Operation code P – Lower female genital tract			
No	6357 (97.5%)	5645 (88.8%)	712 (11.2%)
Yes	160 (2.5%)	141 (88.1%)	19 (11.9%)
Operation code Q – Upper female genital tract			
No	5699 (87.4%)	5052 (88.6%)	647 (11.4%)

Yes	818 (12.6%)	734 (89.7%)	84 (10.3%)
Operation code R – Female genital tract			
No	6378 (97.9%)	5660 (88.7%)	718 (11.3%)
Yes	139 (2.1%)	126 (90.6%)	13 (9.4%)
Operation code S - Skin			
No	5966 (91.5%)	5309 (89.0%)	657 (11.0%)
Yes	551 (8.5%)	477 (86.6%)	74 (13.4%)
Operation code T – Soft tissue			
No	5827 (89.4%)	5198 (89.2%)	629 (10.8%)
Yes	690 (10.6%)	588 (85.2%)	102 (14.8%)
Operation code U – Diagnostic imaging, testing and rehabilitation			
No	6145 (94.3%)	5470 (89.0%)	675 (11.0%)
Yes	372 (5.7%)	316 (84.9%)	56 (15.1%)
Operation code V – Bones and joints of skull and spine			
No	6358 (97.6%)	5647 (88.8%)	711 (11.2%)
Yes	159 (2.4%)	139 (87.4%)	20 (12.6%)
Operation code W – Other bones and joints			
No	5736 (88.0%)	5101 (88.9%)	635 (11.1%)
Yes	781 (12.0%)	685 (87.7%)	96 (12.3%)
Operation code X – Miscellaneous operation			
No	6157 (94.5%)	5480 (89.0%)	677 (11.0%)
Yes	360 (5.5%)	306 (85.0%)	54 (15.0%)
Operation code Y – Methods of operation not elsewhere classifiable			
No	5085 (78.0%)	4552 (89.5%)	533 (10.5%)
Yes	1432 (22.0%)	1234 (86.2%)	198 (13.8%)
Operation code Z – Subsidiary classification			
No	3864 (59.3%)	3460 (89.5%)	404 (10.5%)
Yes	2653 (40.7%)	2326 (87.7%)	327 (12.3%)

SD: Standard Deviation; APC: admitted patient care; CC: critical care; A&E: accident and emergency; OP: outpatients.

Table S4. Associations between analysis variables and HES predictors of non-response at sweep 9 (age 55) in the 1958 British National Child Development Study (NCDS).

	HES predictors of NCDS age 55 non-response									
	1	2	3	4	5	6	7	8	9	10
Father in professional social class at birth	0.01	0.02	0.06	0.001	0.10	0.54	0.16	0.06	0.72	0.07
Cognitive ability at age 7	<0.001	<0.001	0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
No educational qualifications at age 55	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	0.21	0.04
Single and never married at age 55	0.25	0.28	0.008	0.73	0.004	0.63	0.92	0.33	0.61	0.80

p-values from Wald tests of parameter(s) in logistic regression models. Colour coding corresponds to the magnitude of the p-value, from green (0) to red (1).

1. Number of A&E attendances
2. Proportion of OP appointments missed
3. Treatment by Adult Mental Illness
4. ICD Chapter IV: Endocrine, nutritional and metabolic diseases
5. ICD Chapter V: Mental and behavioural disorders
6. ICD Chapter VI: Diseases of the nervous system
7. ICD Chapter X: Diseases of the respiratory system
8. ICD Chapter XVIII: Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified
9. Operation code H: Lower digestive tract
10. Operation code T: Soft tissue

Table S5. Associations between analysis variables and survey predictors of non-response at sweep 9 (age 55) in the 1958 British National Child Development Study (NCDS).

	Survey predictors of NCDS age 55 non-response														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Father in professional social class at birth	<0.001	<0.001	<0.001	<0.001	0.93		<0.001	<0.001	<0.001	<0.001	<0.001	0.001	<0.001	<0.001	0.004
Cognitive ability at age 7	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001		<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
No educational qualifications at age 55	0.37	<0.001	<0.001	<0.001	0.74	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
Single and never married at age 55	<0.001	0.67	0.41	0.12	<0.001	0.001	0.01	0.86	0.01	0.58	0.08	0.13	<0.001	0.85	<0.001

	Survey predictors of NCDS age 55 non-response														
	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
Father in professional social class at birth	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	0.58	<0.001	0.06	<0.001	<0.001	0.03	0.35	0.04	0.001
Cognitive ability at age 7	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	0.01	<0.001	<0.001
No educational qualifications at age 55	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	0.001	0.002	<0.001	<0.001
Single and never married at age 55	0.55	0.12	<0.001	0.71	<0.001	0.75	0.15	<0.001	<0.001	0.01	<0.001	<0.001	<0.001	0.11	<0.001

p-values from Wald tests of parameter(s) in logistic regression models. Colour coding corresponds to the magnitude of the p-value, from green (0) to red (1).

1. Mother's age (Sweep 0)
2. Number of persons per room (Sweep 0)
3. Parity (Sweep 0)
4. Social class of mother's father when she left school (Sweep 0)
5. Sex of child (Sweep 0)
6. Social class of mother's husband (Sweep 0)
7. Dad stayed on at school after minimum age (Sweep 1)
8. Social problems (alcoholism etc.) (Sweep 1)
9. Cognitive ability summary (Sweep 1)
10. Ever breastfed (Sweep 1)
11. Cognitive ability summary (Sweep 2)
12. Conduct problems (Sweep 2)
13. Child receiving help at school – backwardness (Sweep 3)
14. Child's school attendance (Sweep 3)
15. How long since child drank alcohol (Sweep 3)
16. Test 2 – mathematics comprehension (Sweep 3)
17. Conduct problems (Sweep 3)

18. Legal marital status (Sweep 4)
19. Voted in 1979 general election (Sweep 4)
20. Telephone in home (Sweep 5)
21. How much physical effort in job (Sweep 5)
22. Voted in 1987 general election (Sweep 5)
23. Housing tenure (Sweep 5)
24. Social capital score (people turn to for advice, support) (Sweep 5)
25. Membership in organisations (Sweep 6)
26. Self-rated general health (Biomedical sweep)
27. Marital status - de facto (Sweep 7)
28. Total number of natural children (Sweep 8)
29. Employer provided pension scheme (Sweep 8)
30. Non-response at sweeps 1-8

Note: Results omitted in cases where the analysis variable is derived from (or is identical to) the survey predictor of NCDS age 55 non-response.

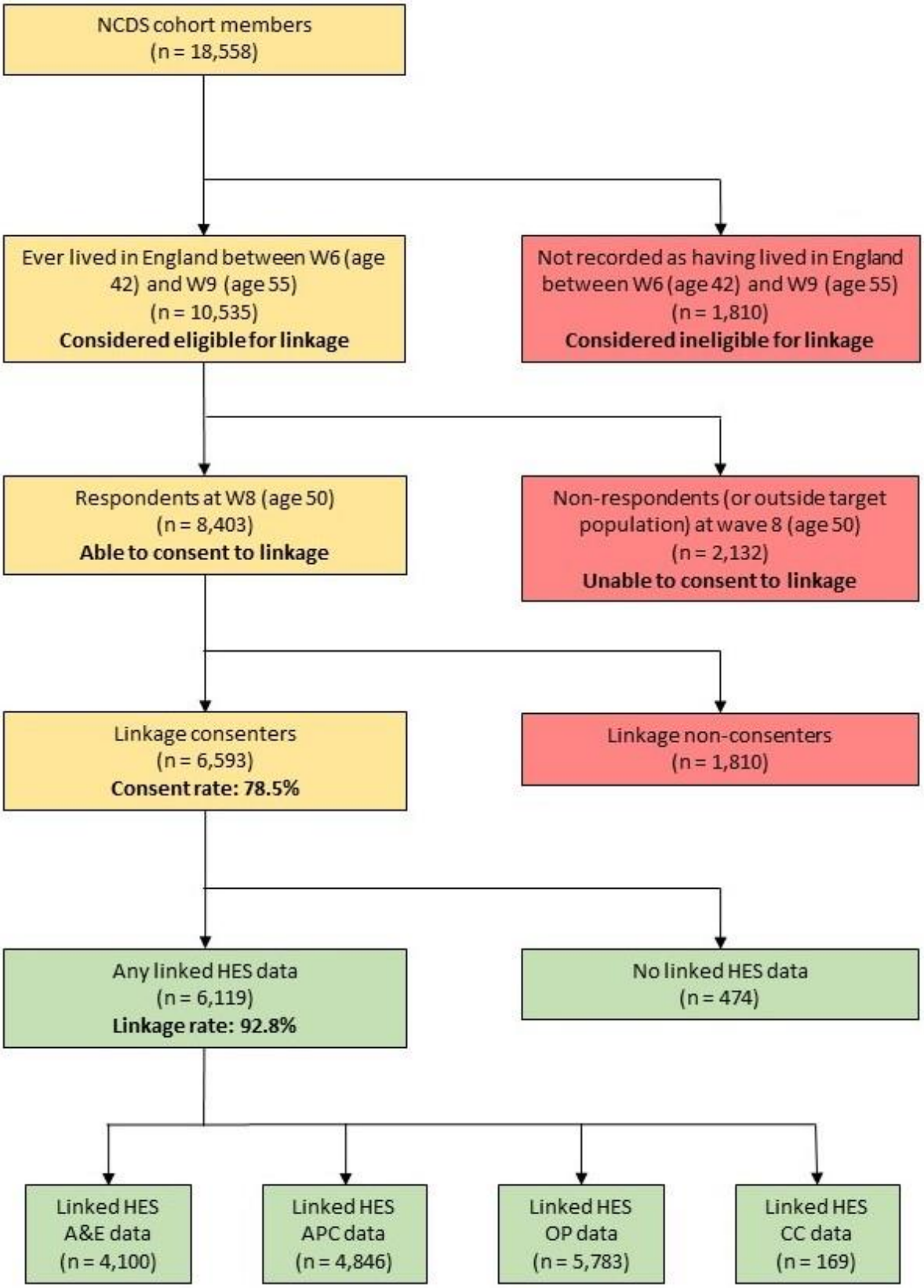

Figure S1. Flow diagram showing 1958 British National Child Development Study-Hospital Episode Statistics data linkage and data availability.
We consider cohort members to be eligible for linkage with HES data if they reported living in England at any one or more waves between waves 6 (2000, age 42) and 9 (2013, age 55),

corresponding to the period of HES data availability (1997 onwards). Of the 10,535 cohort members meeting this definition of linkage eligibility, 8,403 (79.8%) responded at wave 8, with 6,593 (78.5% of respondents) providing consent for linkage. Among these linkage consenters, 6,119 had linked data from one or more of the HES datasets, giving a linkage rate of 92.8%.
APC: admitted patient care; CC: critical care; A&E: accident and emergency; OP: outpatients.

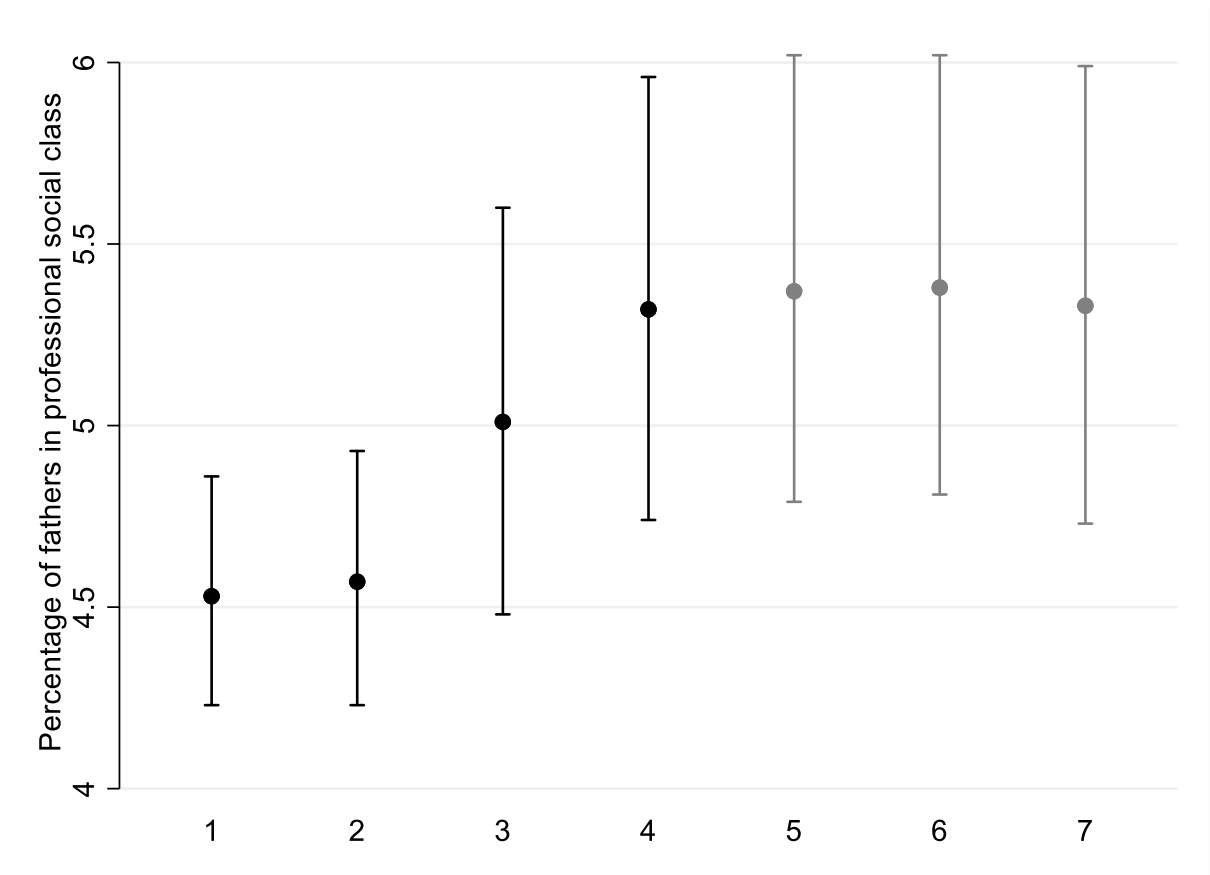

Figure S2. Percentage (95% confidence interval) of fathers in professional social class at birth in the National Child Development Study before and after handling missing data (Analysis A).
Analysis 1: Distribution using all available data (n = 16,458).
Analysis 2: Distribution restricted to wave 9 target population (alive and still living in GB) (n = 13,880).
Analysis 3: Distribution restricted to wave 9 target population and HES linkage consenters who were eligible for linkage (lived in England for at least one wave been W6 and W9) (n = 5,867).
Analysis 4: Distribution restricted to wave 9 respondents within HES linkage consenters who were eligible for linkage (n = 5,226).
Analysis 5: MI analysis using selected survey predictors of non-response as auxiliary variables, restricted to wave 9 target population, HES linkage consenters who were eligible for linkage and those non-missing for the variable of interest, using information on the variable of interest from wave 9 respondents only (n = 5,867).
Analysis 6: MI analysis using both selected HES predictors of non-response and selected survey predictors of non-response as auxiliary variables, restricted to wave 9 target population, HES linkage consenters who were eligible for linkage and those non-missing for the variable of interest, using information on the variable of interest from wave 9 respondents only (n = 5,867).
Analysis 7: MI analysis using selected HES predictors of non-response as auxiliary variables, restricted to wave 9 target population, HES linkage consenters who were eligible for linkage and those non-missing for the variable of interest, using information on the variable of interest from wave 9 respondents only (n = 5,867).

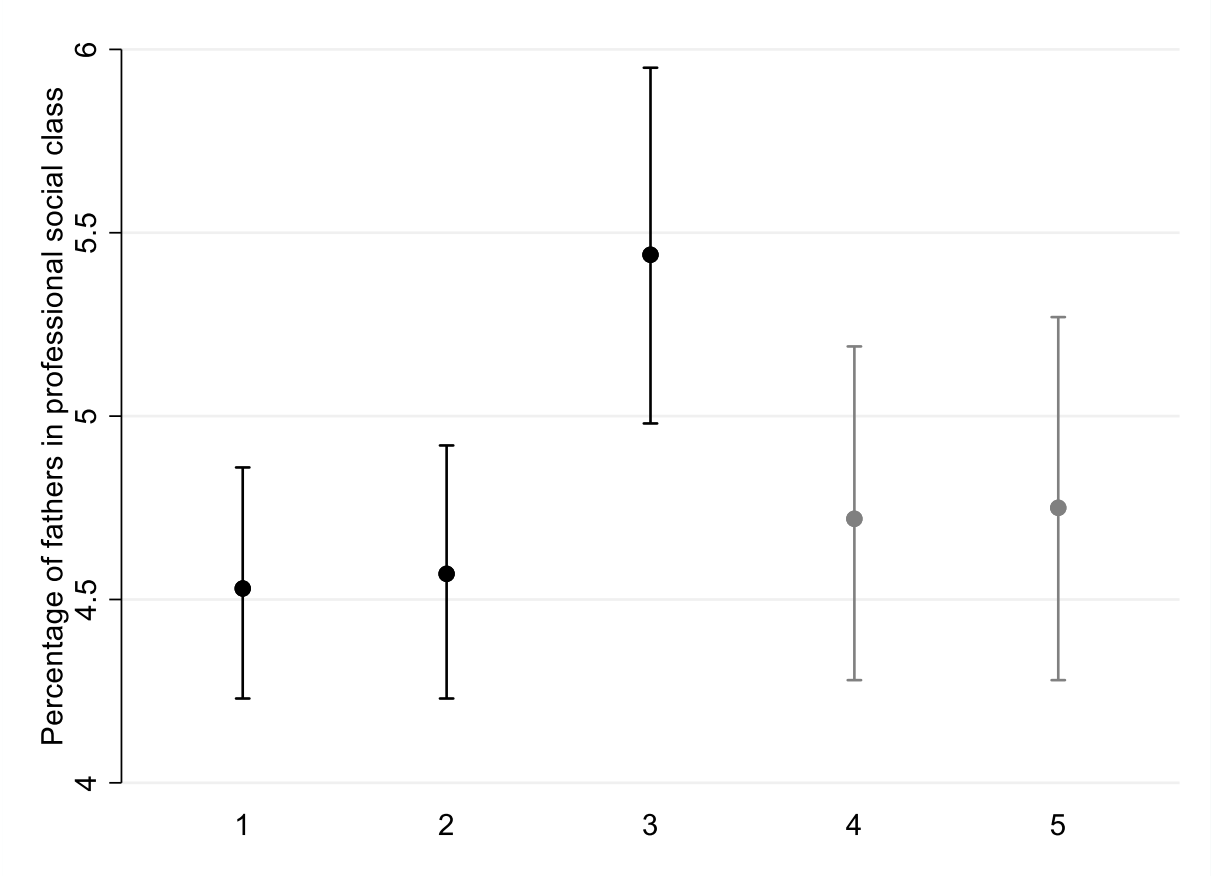

Figure S3. Percentage (95% confidence interval) of fathers in professional social class at birth in the National Child Development Study before and after handling missing data (Analysis B).
Analysis 1: Distribution using all available data (n = 16,458).
Analysis 2: Distribution restricted to wave 9 target population (alive and still living in GB) (n = 13,880).
Analysis 3: Distribution restricted to wave 9 respondents (n = 8,284).
Analysis 4: MI analysis using selected survey predictors of non-response as auxiliary variables, restricted to wave 9 target population and those non-missing for the variable of interest, using information on the variable of interest from wave 9 respondents only (n = 13,880).
Analysis 5: MI analysis using both selected HES predictors of non-response and selected survey predictors of non-response as auxiliary variables, restricted to wave 9 target population and those non-missing for the variable of interest, using information on the variable of interest from wave 9 respondents only (n = 13,880).

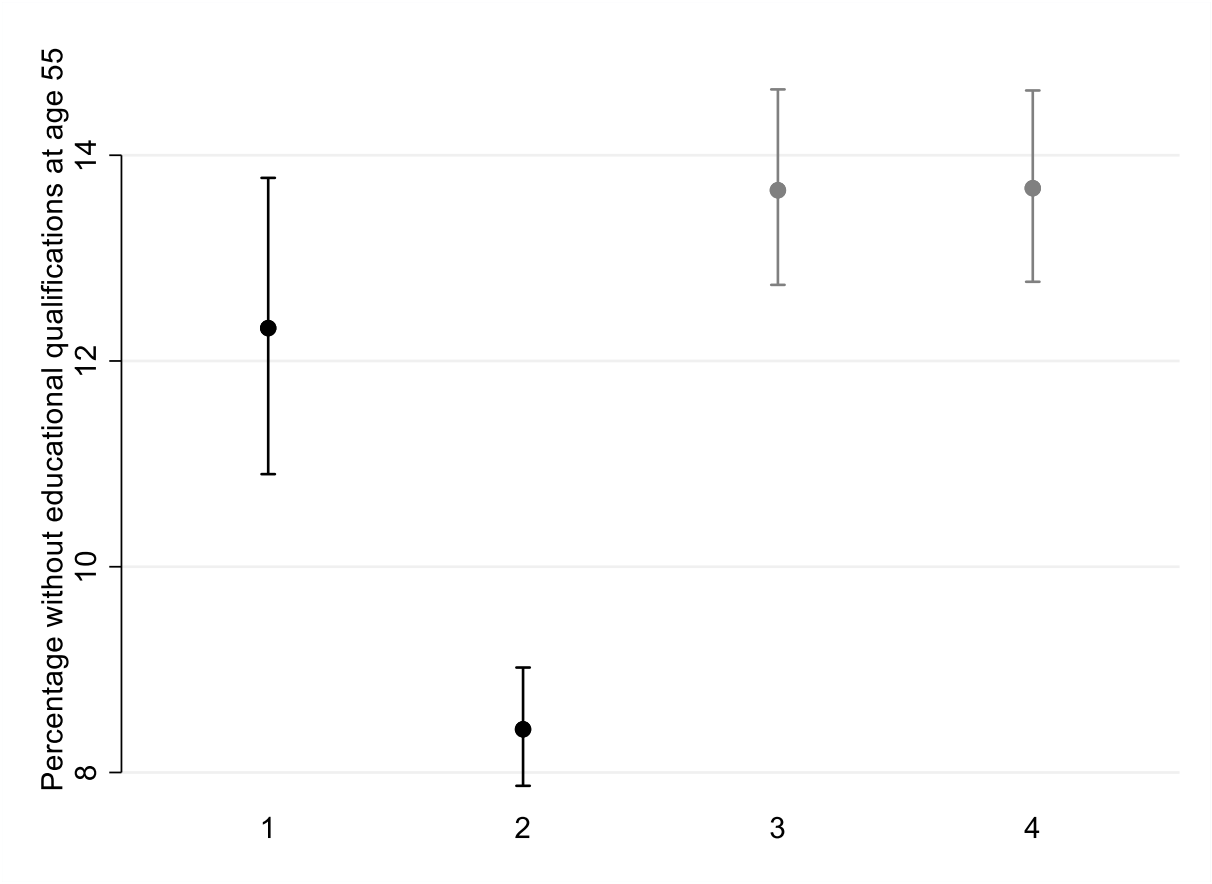

Figure S4. Percentage (95% confidence interval) of cohort members without any educational qualifications at age 55 in the National Child Development Study before and after handling missing data (Analysis B).
Analysis 1: Distribution using population benchmark data from APS (n = 1,935).
Analysis 2: Distribution among wave 9 respondents (which in this case is all available information) (n = 8,952).
Analysis 3: MI analysis using selected survey predictors of non-response as auxiliary variables, restricted to wave 9 target population (n = 15,613).
Analysis 4: MI analysis using both selected HES predictors of non-response and selected survey predictors of non-response as auxiliary variables, restricted to wave 9 target population (n = 15,613).

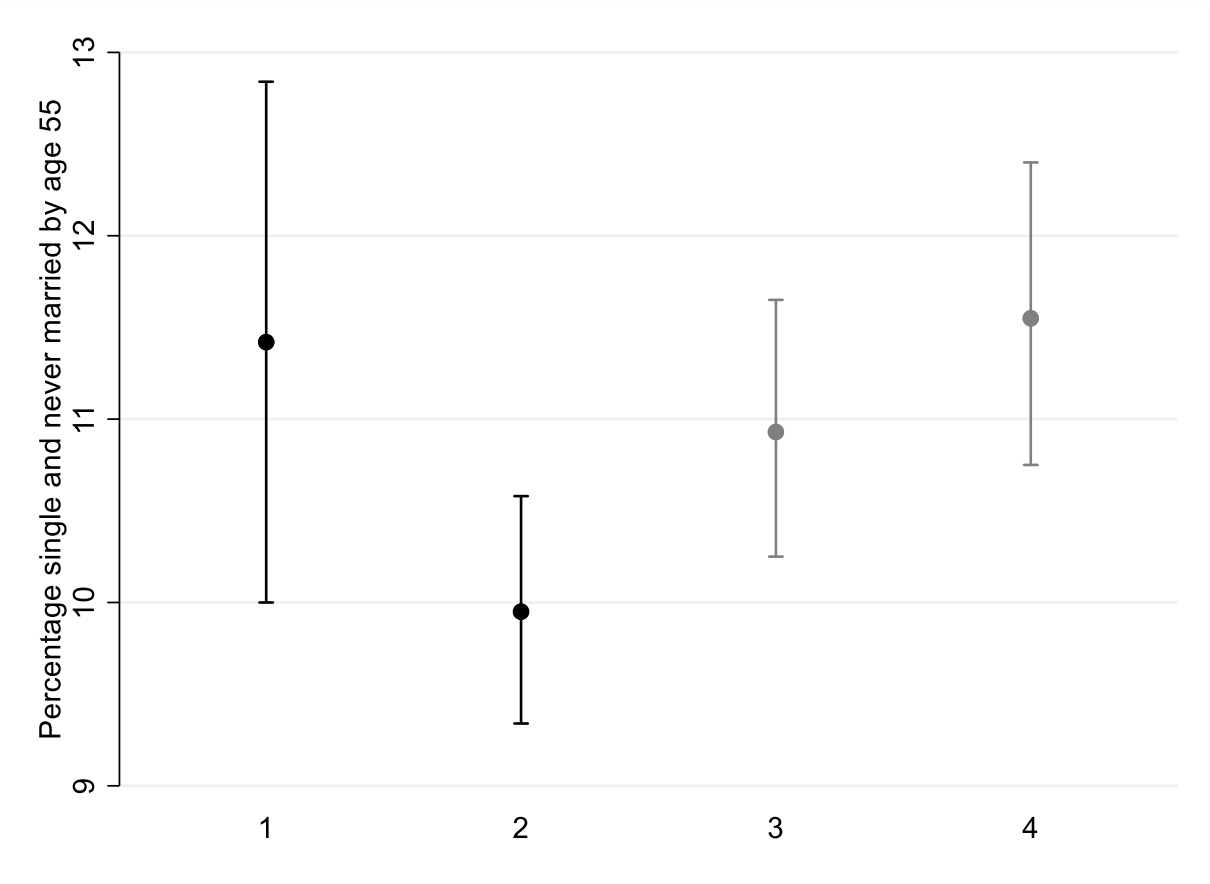

Figure S5. Percentage (95% confidence interval) of cohort members who are single and never married at age 55 in the National Child Development Study before and after handling missing data (Analysis B).
Analysis 1: Distribution using population benchmark data from APS (n = 1,937).
Analysis 2: Distribution among wave 9 respondents (which in this case is all available information) (n = 9,130).
Analysis 3: MI analysis using selected survey predictors of non-response as auxiliary variables, restricted to wave 9 target population (n = 15,613).
Analysis 4: MI analysis using both selected HES predictors of non-response and selected survey predictors of non-response as auxiliary variables, restricted to wave 9 target population (n = 15,613).