Handling missing data in the 1970 British Cohort Study

Richard Silverwood, Michalis Katsoulis & Brian Dodgeon

CENTRE FOR LONGITUDINAL STUDIES

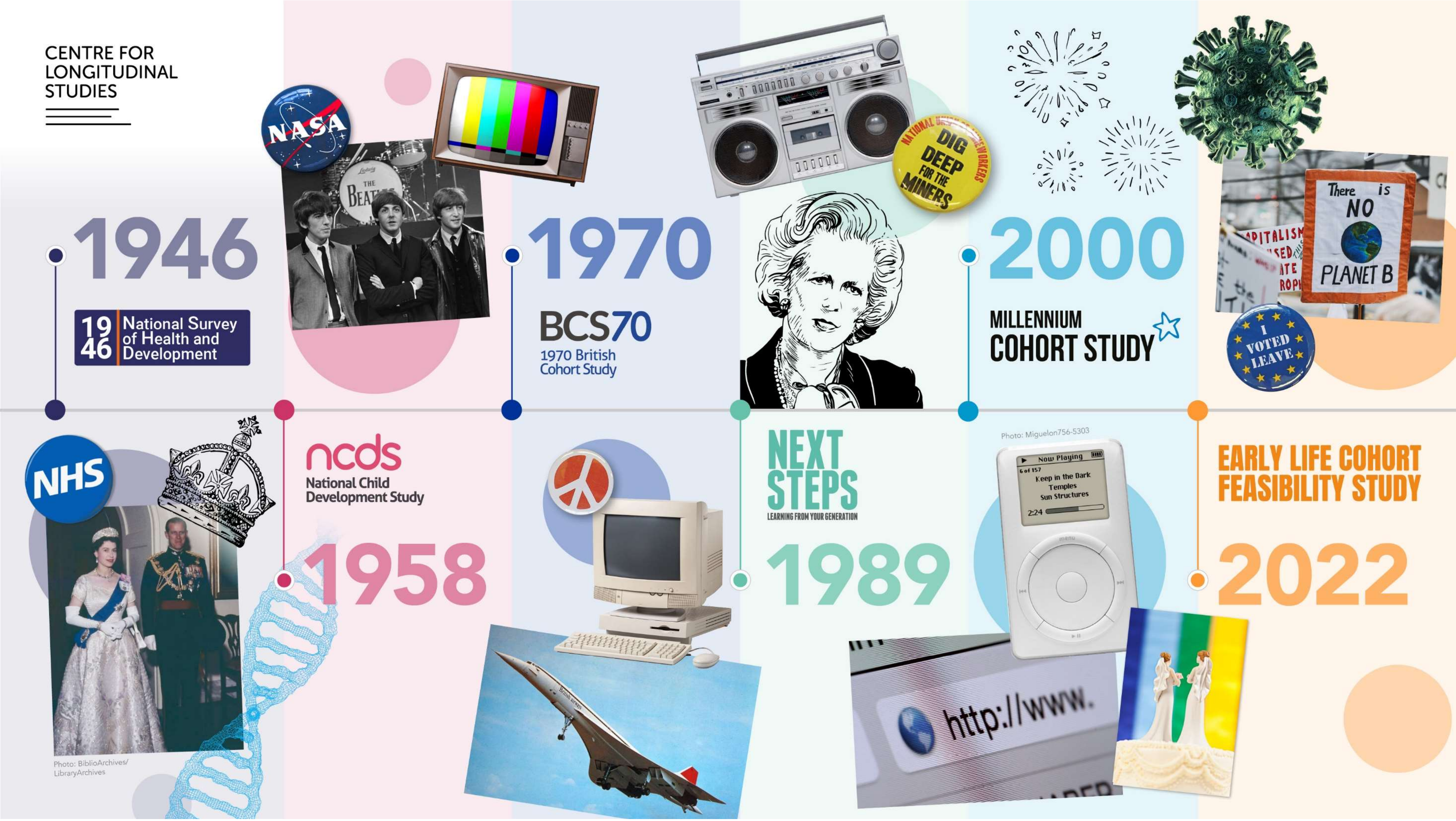
6 June 2024

Economic
and Social
Research Council

CENTRE FOR
LONGITUDINAL
STUDIES

Introduction to BCS70

CENTRE FOR
LONGITUDINAL
STUDIES



1946

19 46 National Survey of Health and Development

1958

ncds
National Child Development Study

1970

BCS70
1970 British Cohort Study

1989

NEXT STEPS
LEARNING FROM YOUR GENERATION

2000

MILLENNIUM COHORT STUDY

2022

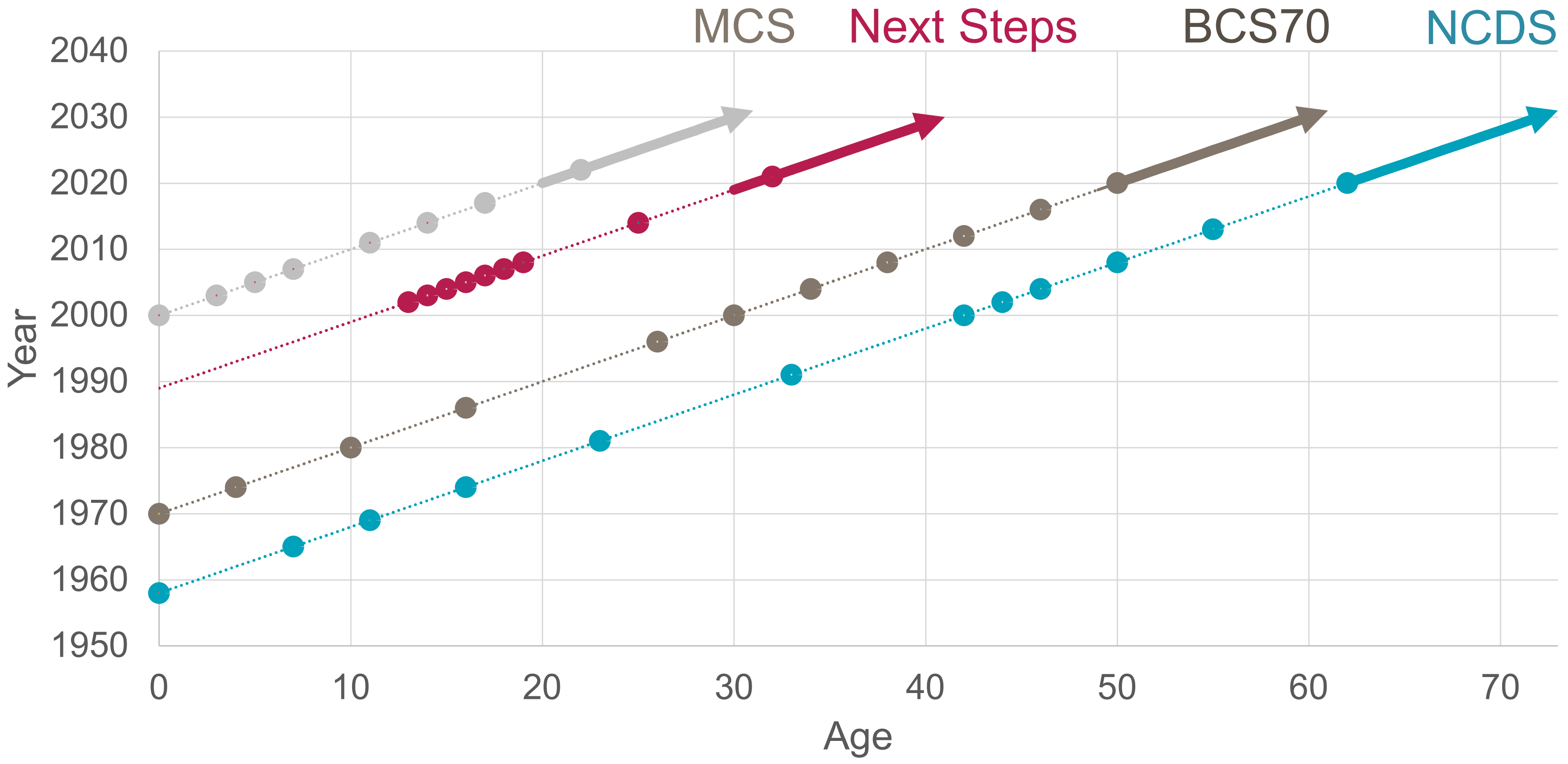
EARLY LIFE COHORT FEASIBILITY STUDY

Photo: BiblioArchives/ LibraryArchives

Photo: Miguelon756-5303

CLS core study timelines

CENTRE FOR
LONGITUDINAL
STUDIES



1970 British Cohort Study (BCS70)

CENTRE FOR LONGITUDINAL STUDIES

- Longitudinal birth cohort study of people born in England, Wales, Scotland and Northern Ireland in one week in 1970.
- Initial N = 17,196, though NI members not followed up after birth (unless moved to mainland).
- Boost samples through childhood.
- Collected information on health, physical, educational and social development, and economic circumstances, among other factors.
- PI: George Ploubidis.

BCS70: Respondents, instruments and response

CENTRE FOR
LONGITUDINAL
STUDIES

	1970	1975	1980	1986	1996	2000	2004	2008	2012	2016
	Birth	5	10	16	26	30	34	38	42	46
main respondent	mother	parents	subject/parents	subject/parents	subject	subject	subject	subject	subject	subject
secondary respondent	medical	medical	medical/school	medical/school			children			medical
survey instruments		cognitive assessments	cognitive assessments	cognitive assessments and diaries	16-page postal questionnaire		cognitive assessments	telephone survey	vocabulary test	cognitive assessments
linked data									Consent to data linkage	
response	17,196	13,135	14,875	11,622	9,003	11,261	9,665	8,874	9,841	8,581

BCS70: Topics covered by life stage

Birth

- Family
- Parental employment
- Obstetric history
- Smoking in pregnancy
- Pregnancy (problems, antenatal care)
- Labour (length, pain relief, problems)
- Birth (problems, sex, weight, length)

School years

- Family
- Parental employment
- Financial circumstances
- Housing
- Health
- Behaviour
- School
- Views and expectations
- Attainment

Adult

- Family (partners, children)
- Employment
- Income
- Housing
- Health
- Health-related behaviour
- Courses and qualifications
- Basic skills
- Cognitive ability
- Views and expectations

Age 51 (available November 2024)

CENTRE FOR
LONGITUDINAL
STUDIES

Employment and income

- Occupation
- Income
- Partner’s employment and income
- Impact of COVID-19
- Benefits
- Pensions
- Debts
- Intergenerational transfers

Physical health and health behaviours

- General
- Longstanding illness
- Health conditions
- COVID tests, symptoms, long COVID
- Height/weight
- Exercise
- Diet
- Drinking and smoking

Cognitive skills and processes

- Immediate and delayed recall
- Animal naming
- Letter cancellation
- National Audit Reading Test (NART)

Mental health and well-being

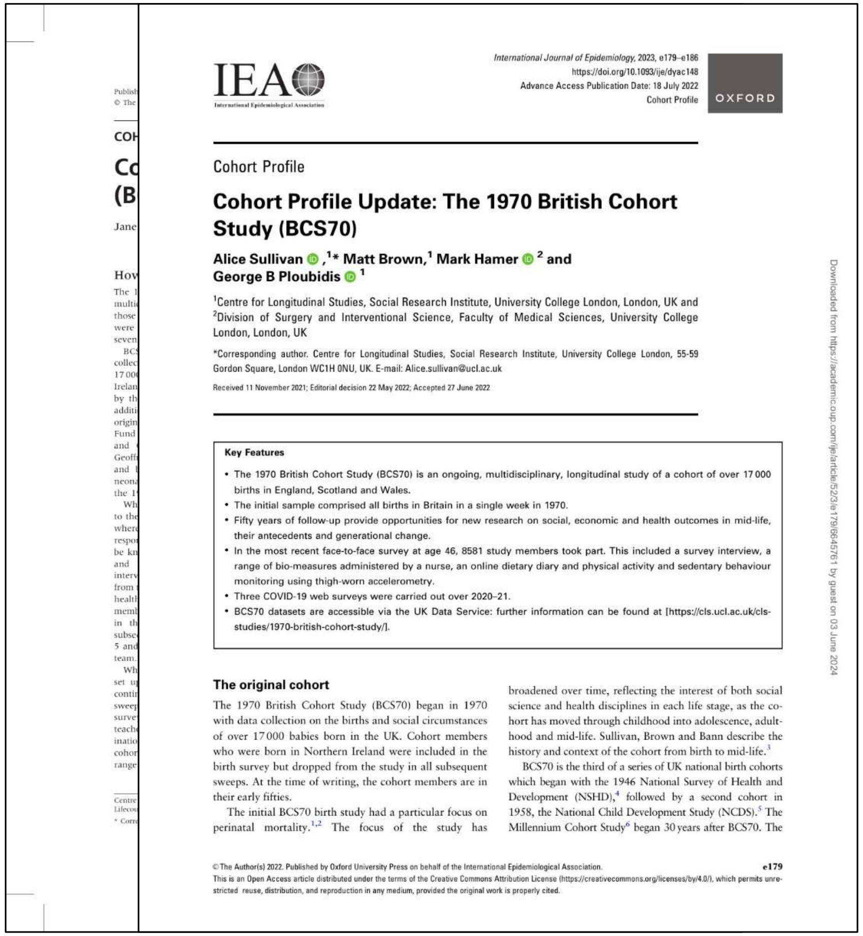
- Psychological distress
- Mental well-being
- Life satisfaction

Family and relationships

- Grandchildren
- CM’s and partners parents – caring responsibilities
- Social contact
- Quality of relationships
- Menstruation
- Fertility

Activities, attitudes and values

Cohort profiles



International Journal of Epidemiology, 2023, e179–e186
https://doi.org/10.1093/ije/dyac148
Advance Access Publication Date: 18 July 2022
Cohort Profile

Cohort Profile

Cohort Profile Update: The 1970 British Cohort Study (BCS70)

Alice Sullivan,[1]* **Matt Brown**,[1] **Mark Hamer**[2] **and George B Ploubidis**[1]

[1]Centre for Longitudinal Studies, Social Research Institute, University College London, London, UK and [2]Division of Surgery and Interventional Science, Faculty of Medical Sciences, University College London, London, UK

*Corresponding author. Centre for Longitudinal Studies, Social Research Institute, University College London, 55-59 Gordon Square, London WC1H 0NU, UK. E-mail: Alice.sullivan@ucl.ac.uk

Received 11 November 2021; Editorial decision 22 May 2022; Accepted 27 June 2022

Key Features

- The 1970 British Cohort Study (BCS70) is an ongoing, multidisciplinary, longitudinal study of a cohort of over 17 000 births in England, Scotland and Wales.
- The initial sample comprised all births in Britain in a single week in 1970.
- Fifty years of follow-up provide opportunities for new research on social, economic and health outcomes in mid-life, their antecedents and generational change.
- In the most recent face-to-face survey at age 46, 8581 study members took part. This included a survey interview, a range of bio-measures administered by a nurse, an online dietary diary and physical activity and sedentary behaviour monitoring using thigh-worn accelerometry.
- Three COVID-19 web surveys were carried out over 2020–21.
- BCS70 datasets are accessible via the UK Data Service: further information can be found at [https://cls.ucl.ac.uk/cls-studies/1970-british-cohort-study/].

The original cohort

The 1970 British Cohort Study (BCS70) began in 1970 with data collection on the births and social circumstances of over 17 000 babies born in the UK. Cohort members who were born in Northern Ireland were included in the birth survey but dropped from the study in all subsequent sweeps. At the time of writing, the cohort members are in their early fifties.

The initial BCS70 birth study had a particular focus on perinatal mortality.[1,2] The focus of the study has broadened over time, reflecting the interest of both social science and health disciplines in each life stage, as the cohort has moved through childhood into adolescence, adulthood and mid-life. Sullivan, Brown and Bann describe the history and context of the cohort from birth to mid-life.[3]

BCS70 is the third of a series of UK national birth cohorts which began with the 1946 National Survey of Health and Development (NSHD),[4] followed by a second cohort in 1958, the National Child Development Study (NCDS).[5] The Millennium Cohort Study[6] began 30 years after BCS70. The

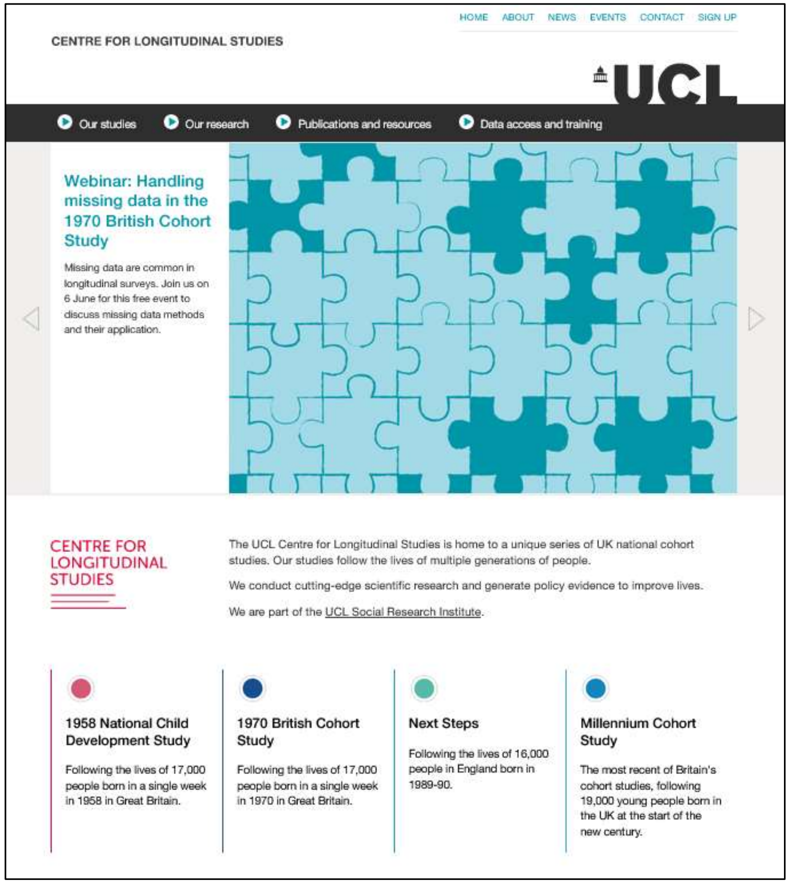
Downloaded from https://academic.oup.com/ije/article/52/3/e179/6645761 by guest on 03 June 2024

© The Author(s) 2022. Published by Oxford University Press on behalf of the International Epidemiological Association.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Elliott J, Shepherd P. Cohort Profile: 1970 British Birth Cohort (BCS70). International Journal of Epidemiology. 2006; 35(4): 836-43.

Sullivan A, Brown M, Hamer M, Ploubidis G. Cohort Profile Update: The 1970 British Cohort Study (BCS70). International Journal of Epidemiology. 2023; 52(3): e179-e86.

Accessing more information/data



Webinar: Handling missing data in the 1970 British Cohort Study

Missing data are common in longitudinal surveys. Join us on 6 June for this free event to discuss missing data methods and their application.

The UCL Centre for Longitudinal Studies is home to a unique series of UK national cohort studies. Our studies follow the lives of multiple generations of people.

We conduct cutting-edge scientific research and generate policy evidence to improve lives.

We are part of the UCL Social Research Institute.

1958 National Child Development Study

Following the lives of 17,000 people born in a single week in 1958 in Great Britain.

1970 British Cohort Study

Following the lives of 17,000 people born in a single week in 1970 in Great Britain.

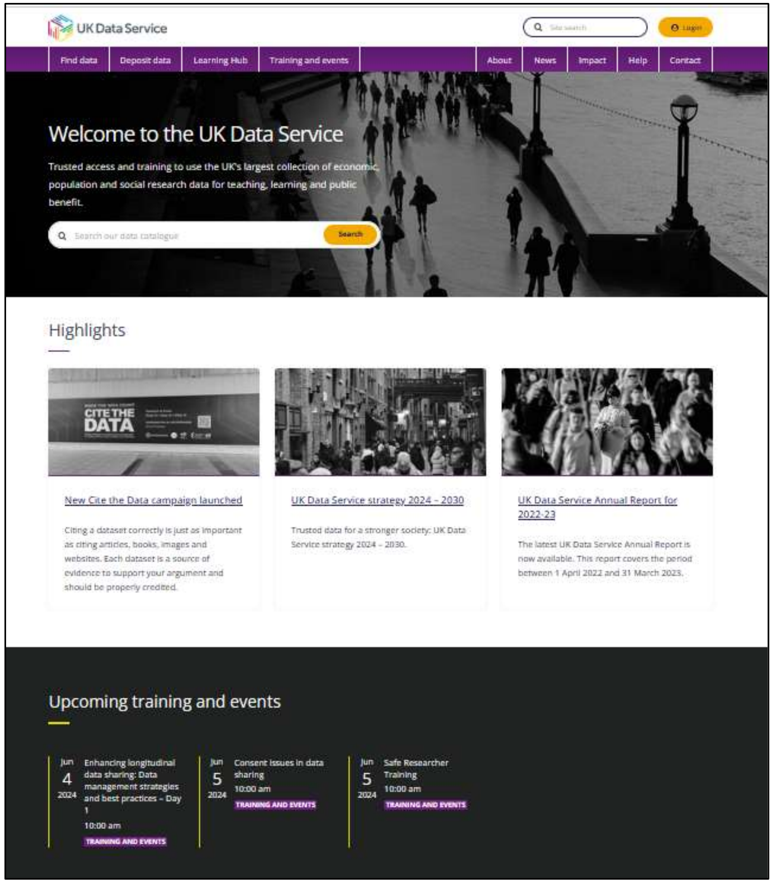
Next Steps

Following the lives of 16,000 people in England born in 1989-90.

Millennium Cohort Study

The most recent of Britain's cohort studies, following 19,000 young people born in the UK at the start of the new century.

https://cls.ucl.ac.uk



Welcome to the UK Data Service

Trusted access and training to use the UK's largest collection of economic, population and social research data for teaching, learning and public benefit.

Highlights

New Cite the Data campaign launched

Citing a dataset correctly is just as important as citing articles, books, images and websites. Each dataset is a source of evidence to support your argument and should be properly credited.

UK Data Service strategy 2024 – 2030

Trusted data for a stronger society: UK Data Service strategy 2024 – 2030.

UK Data Service Annual Report for 2022-23

The latest UK Data Service Annual Report is now available. This report covers the period between 1 April 2022 and 31 March 2023.

Upcoming training and events

Jun 4 2024 Enhancing longitudinal data sharing: Data management strategies and best practices – Day 1
10:00 am
TRAINING AND EVENTS

Jun 5 2024 Consent issues in data sharing
10:00 am
TRAINING AND EVENTS

Jun 5 2024 Safe Researcher Training
10:00 am
TRAINING AND EVENTS

https://ukdataservice.ac.uk/

CENTRE FOR
LONGITUDINAL
STUDIES

CLS missing data strategy

Missing data

- Non-response inevitable in longitudinal surveys.
- Reduced analysis sample size → reduced efficiency.
- Respondents often systematically different from non-respondents → threat to representativeness & potential for bias.
- Well known (principled) methods for handling missing data include multiple imputation, inverse probability weighting and full information maximum likelihood.
- These rely on the assumption that the data are missing at random (MAR) = given the observed values, missingness does not depend on unobserved values.

Multiple imputation (MI)

- Specify an appropriate imputation model and create a series of imputed datasets.
- Each imputed dataset is analysed using the substantive model and results combined using standard rules.
- Need to include all variables in the substantive model in the imputation model.
- Can also include “auxiliary variables”: variables associated with the underlying values of the variable(s) subject to missingness, particularly those also associated with the probability of missingness.

CLS Missing Data Strategy

- Most analyses employing such MAR methods rely on a largely arbitrary selection of variables used as predictors of missingness.
- We aim to maximise the plausibility of the MAR assumption by optimising the set of such variables used in analyses.
- We use systematic, data-driven approaches to identify variables that are associated with non-response at each wave.
- These can then be considered for inclusion as auxiliary variables.
- This allows us to capitalise on the rich data cohort members have provided over the years/decades.

Why the focus on wave non-response?

- Wave non-response is the main driver of missing data in analyses of CLS studies. Item non-response less of an issue.
- Much of the wave non-response is due to attrition.
- For longitudinal analyses, wave non-response at the most recent sweep is therefore usually the biggest contributor to missingness.
- Can identify predictors of wave non-response at cohort (rather than analysis) level – pragmatic approach.
- In analyses in which item non-response is more prevalent, this may need additional consideration.

1958 National Child Development Study

CENTRE FOR
LONGITUDINAL
STUDIES

Longitudinal and Life Course Studies • vol 13 • no 2 • 335–341
© Authors 2022 • Paper received October 2021/Published online Feburary 2022
Online ISSN 1757-9597 • https://doi.org/10.1332/175795921X16428748347208

This article is distributed under the terms of the Creative Commons Attribution 4.0 license (http://creativecommons.org/licenses/by/4.0/) which permits adaptation, alteration, reproduction and distribution without further permission provided the original work is attributed. The derivative works do not need to be licensed on the same terms.

CORRESPONDENCE

Letter to the editor: Don't forget survey data: 'healthy cohorts' are 'real-world' relevant if missing data are handled appropriately

Richard J. Silverwood, r.silverwood@ucl.ac.uk
Alissa Goodman, alissa.goodman@ucl.ac.uk
George B. Ploubidis, g.ploubidis@ucl.ac.uk
University College London, UK

To cite this article: Silverwood, R., Goodman, A. and Ploubidis, G. (2022) 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*, vol 13, no 2, 335–341, DOI: 10.1332/175795921X16428748347208

Dear Professor Joshi,

We write to you regarding the published article 'Are "healthy cohorts" real-world relevant? Comparing the National Child Development Study (NCDS) with the ONS Longitudinal Study (LS)' by Archer et al (2020). The authors report that NCDS is unrepresentative of age-matched LS respondents, but that despite differences in sample characteristics, longitudinal associations were similar in the NCDS and LS samples. They attribute the discrepancy between NCDS and LS to a 'healthy cohort' effect and propose that creating non-response weights from administrative data should be used. While we agree with Archer et al that administrative data have the potential to inform missing data analyses in longitudinal surveys, the authors omit to mention that even without administrative data there are already methods available to researchers to restore sample representativeness using survey information alone that have been shown to be highly effective.

To demonstrate the effectiveness of using survey information – without augmentation by administrative data – in restoring sample representativeness in NCDS with respect to the LS, we present Table 1 from their manuscript, with additional columns from our own analyses. We accounted for non-response at age 46 and 55 with multiple imputation (MI), using chained equations (Azur et al, 2011; White et al, 2011; Harel et al, 2018) to generate 50 imputed datasets.[1] The imputation phase included 'auxiliary variables' (Carpenter and Kenward 2012) from earlier sweeps of

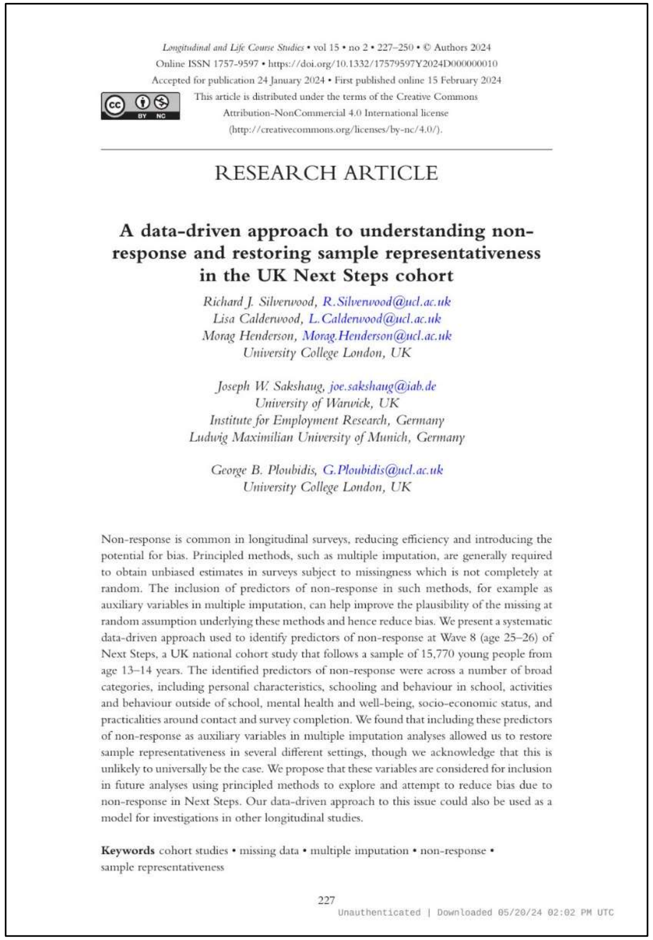
335

Brought to you by University College London | Unauthenticated | Downloaded 06/19/23 03:29 PM UTC

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): 335-341

Mostafa T, Narayanan M, Pongiglione B, Dodgeon B, Goodman A, Silverwood RJ, Ploubidis GB. **Missing at random assumption made more plausible: evidence from the 1958 British birth cohort.** Journal of Clinical Epidemiology 2021; 136: 44-54.

Next Steps



Longitudinal and Life Course Studies • vol 15 • no 2 • 227–250 • © Authors 2024
Online ISSN 1757-9597 • https://doi.org/10.1332/17579597Y2024D000000010
Accepted for publication 24 January 2024 • First published online 15 February 2024

This article is distributed under the terms of the Creative Commons
Attribution-NonCommercial 4.0 International license
(http://creativecommons.org/licenses/by-nc/4.0/).

RESEARCH ARTICLE

A data-driven approach to understanding non-response and restoring sample representativeness in the UK Next Steps cohort

Richard J. Silverwood, R.Silverwood@ucl.ac.uk
Lisa Calderwood, L.Calderwood@ucl.ac.uk
Morag Henderson, Morag.Henderson@ucl.ac.uk
University College London, UK

Joseph W. Sakshaug, joe.sakshaug@iab.de
University of Warwick, UK
Institute for Employment Research, Germany
Ludwig Maximilian University of Munich, Germany

George B. Ploubidis, G.Ploubidis@ucl.ac.uk
University College London, UK

Non-response is common in longitudinal surveys, reducing efficiency and introducing the potential for bias. Principled methods, such as multiple imputation, are generally required to obtain unbiased estimates in surveys subject to missingness which is not completely at random. The inclusion of predictors of non-response in such methods, for example as auxiliary variables in multiple imputation, can help improve the plausibility of the missing at random assumption underlying these methods and hence reduce bias. We present a systematic data-driven approach used to identify predictors of non-response at Wave 8 (age 25–26) of Next Steps, a UK national cohort study that follows a sample of 15,770 young people from age 13–14 years. The identified predictors of non-response were across a number of broad categories, including personal characteristics, schooling and behaviour in school, activities and behaviour outside of school, mental health and well-being, socio-economic status, and practicalities around contact and survey completion. We found that including these predictors of non-response as auxiliary variables in multiple imputation analyses allowed us to restore sample representativeness in several different settings, though we acknowledge that this is unlikely to universally be the case. We propose that these variables are considered for inclusion in future analyses using principled methods to explore and attempt to reduce bias due to non-response in Next Steps. Our data-driven approach to this issue could also be used as a model for investigations in other longitudinal studies.

Keywords cohort studies • missing data • multiple imputation • non-response • sample representativeness

Unauthenticated | Downloaded 05/20/24 02:02 PM UTC

Silverwood RJ, Calderwood L, Henderson M, Sakshaug JW, Ploubidis GB. **A data-driven approach to understanding non-response and restoring sample representativeness in the UK Next Steps cohort**. Longitudinal and Life Course Studies. 2024; 15(2): 227-50.

BCS70

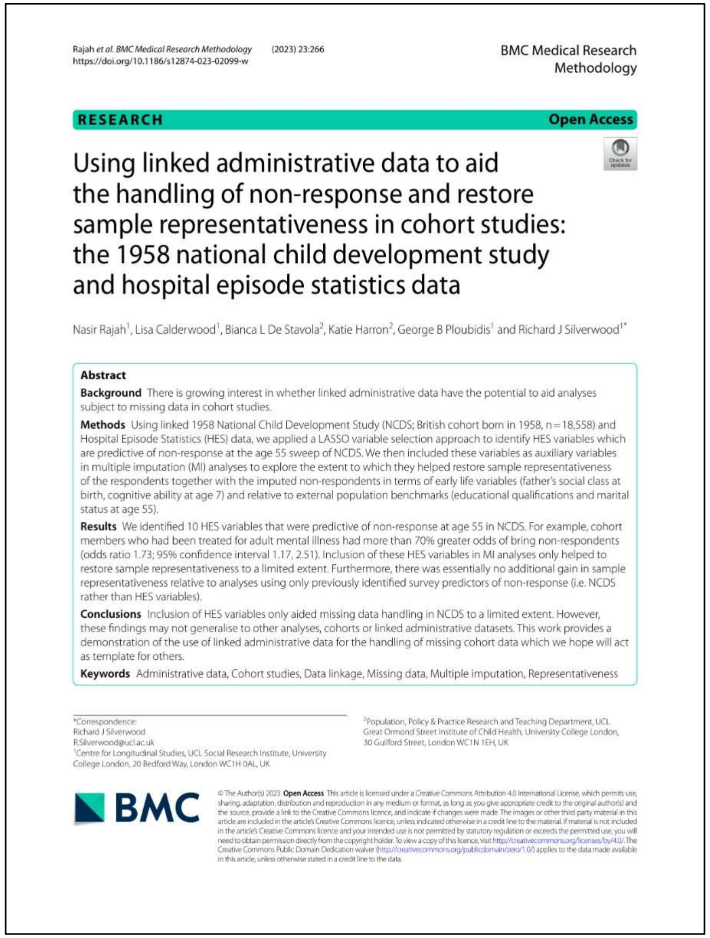
A data driven approach to address missing data in the 1970 British birth cohort

Michail Katsoulis[1], Martina Narayanan[2], Brian Dodgeon[2], George Ploubidis[2], Richard Silverwood[2]

1. MRC unit for Lifelong Health and Ageing, Institute of Cardiovascular Science, UCL
2. Centre for Longitudinal Studies, Institute of Education, UCL

Katsoulis M, Narayanan M, Dodgeon B, Ploubidis G, Silverwood R. **A data driven approach to address missing data in the 1970 British birth cohort**. medRxiv. 2024: 2024.02.01.24302101.

Extending the strategy



Rajah *et al. BMC Medical Research Methodology* (2023) 23:266
https://doi.org/10.1186/s12874-023-02099-w

BMC Medical Research
Methodology

RESEARCH Open Access

Using linked administrative data to aid the handling of non-response and restore sample representativeness in cohort studies: the 1958 national child development study and hospital episode statistics data

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

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, n = 18,558) 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 had more than 70% greater odds of bring non-respondents (odds 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 HES variables only aided missing data handling in NCDS to a limited extent. However, these findings may not generalise to other analyses, cohorts or linked administrative datasets. This work provides a demonstration of the use of linked administrative data for the handling of missing cohort data which we hope will act as template for others.

Keywords Administrative data, Cohort studies, Data linkage, Missing data, Multiple imputation, Representativeness

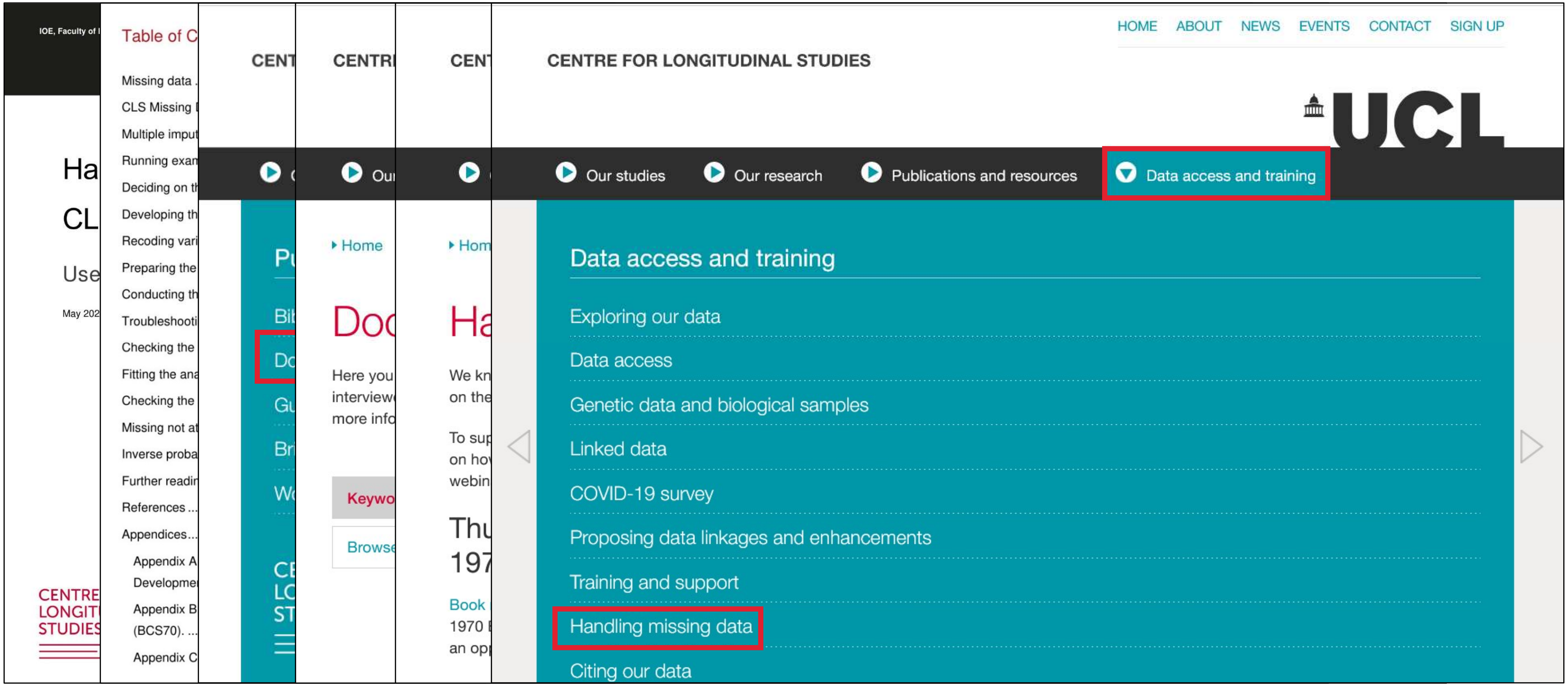
*Correspondence:
Richard J Silverwood
R.Silverwood@ucl.ac.uk
[1]Centre for Longitudinal Studies, UCL Social Research Institute, University College London, 20 Bedford Way, London WC1H 0AL, UK
[2]Population, Policy & Practice Research and Teaching Department, UCL Great Ormond Street Institute of Child Health, University College London, 30 Guilford Street, London WC1N 1EH, UK

© The Author(s) 2023. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

Rajah N, Calderwood L, De Stavola BL, Harron K, Ploubidis GB, Silverwood RJ. **Using linked administrative data to aid the handling of non-response and restore sample representativeness in cohort studies: the 1958 national child development study and hospital episode statistics data**. BMC Medical Research Methodology. 2023; 23(1): 266.

Additional resources

CENTRE FOR
LONGITUDINAL
STUDIES



IOE, Faculty of I

Ha

CL

Use

May 202

CENTRE
LONGIT
STUDIES

Table of C

CENT

Pu

Bit

Do

Gu

Br

W

CE
LO
ST

CENTR

Ou

Home

Doc

Here you
interview
more info

Keywo

Browse

CENT

Hom

Ha

We kn
on the

To sup
on hov
webin

Thu
197

Book
1970 I
an op

CENTRE FOR LONGITUDINAL STUDIES

HOME ABOUT NEWS EVENTS CONTACT SIGN UP

UCL

Our studies | Our research | Publications and resources | Data access and training

Data access and training

- Exploring our data
- Data access
- Genetic data and biological samples
- Linked data
- COVID-19 survey
- Proposing data linkages and enhancements
- Training and support
- Handling missing data
- Citing our data

CENTRE FOR
LONGITUDINAL
STUDIES

Missing data strategy in BCS70

Identifying predictors of non-response in BCS70

- Very similar approach used in BCS70 as in NCDS

- Aim to maximise the plausibility of the MAR assumption using a data driven approach we identify the variables that are associated with non-response at each sweep (and can potentially be used as auxiliary variables)

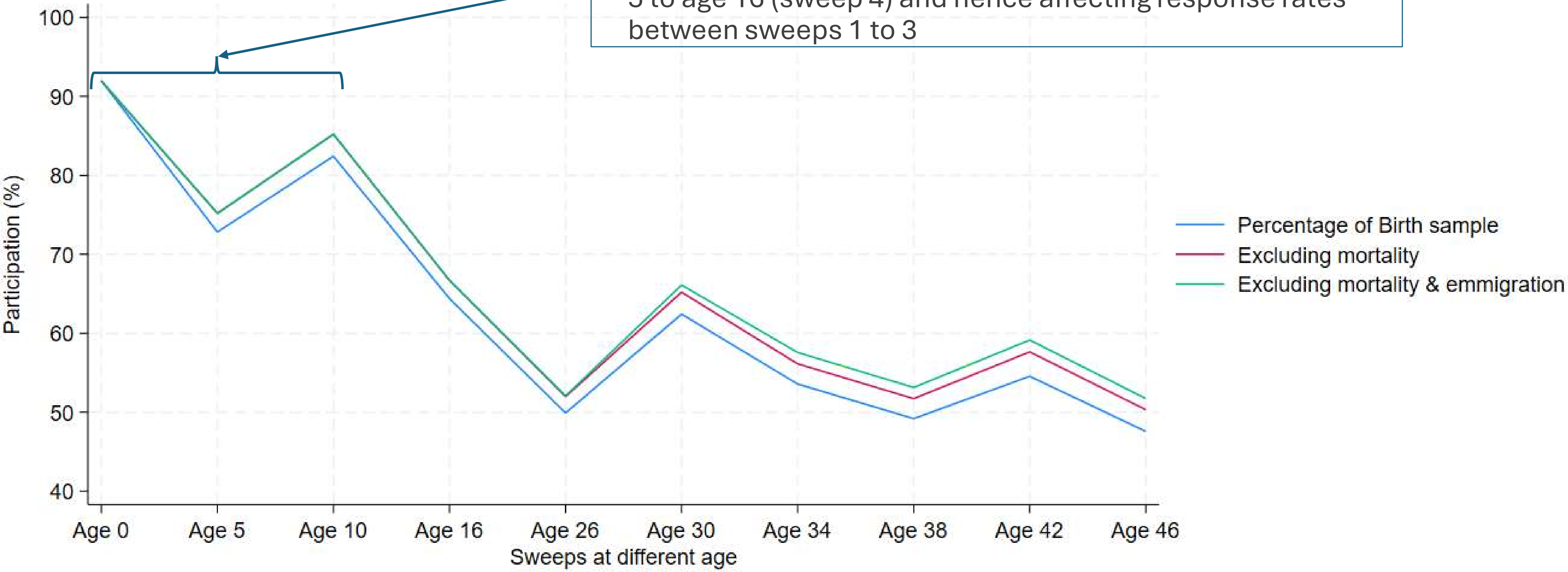
Summary

- We have identified variables which predict non-response at each wave of BCS70.
- These can be used as auxiliary variables in subsequent analyses to increase the plausibility of the MAR assumption.
- Simple test analyses have shown this approach to perform well.
- A straightforward approach, easily implemented in standard software.
- Lists of predictors of non-response available via Handling missing data in the CLS cohort studies (https://cls.ucl.ac.uk/wp-content/uploads/2020/04/Handling-Missing-Data-User-Guide-2024.pdf).
- Will be updated when new waves of data become available.

BCS: Response over time

CENTRE FOR LONGITUDINAL STUDIES

- Individuals who were not resident in the UK in 1970 (here included as not respondents) were added later between age 5 to age 16 (sweep 4) and hence affecting response rates between sweeps 1 to 3



Identifying predictors of non-response in BCS70

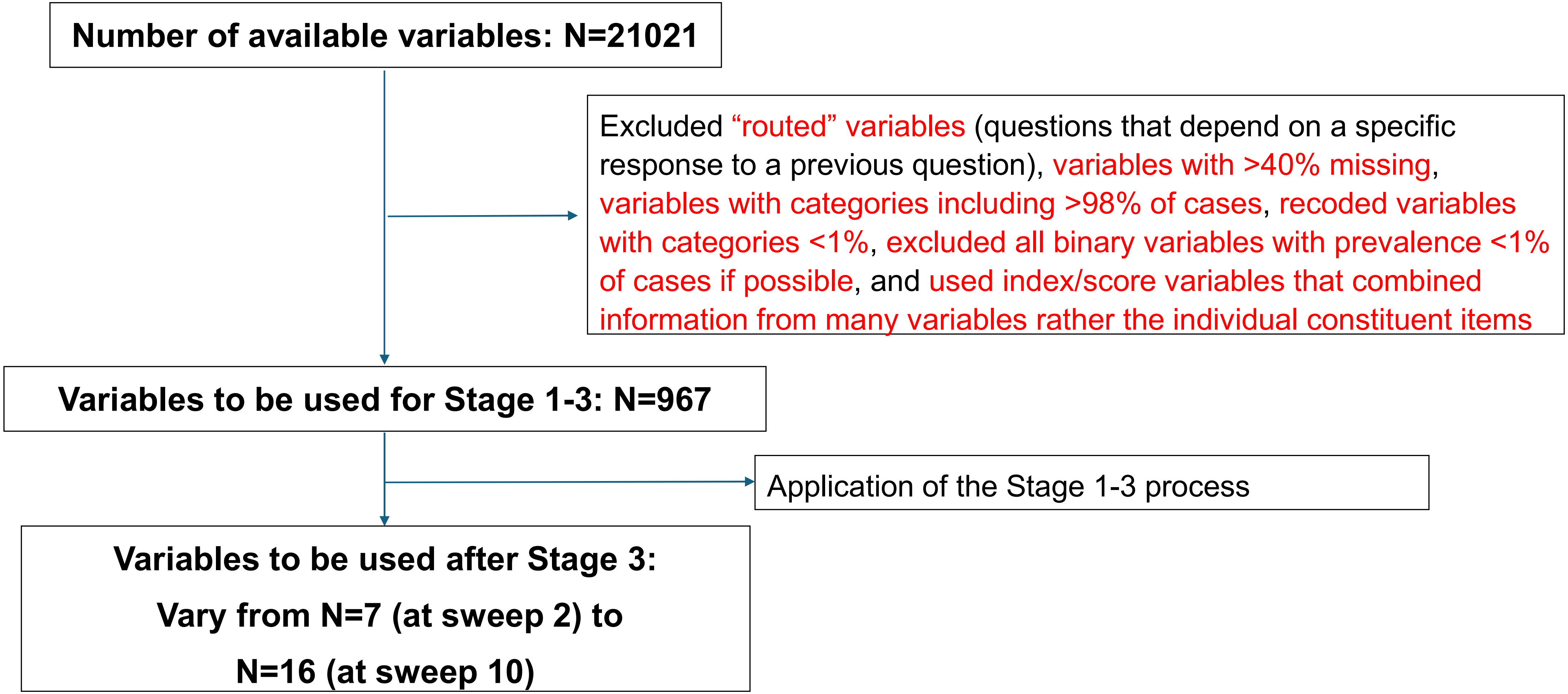
- ~20000 variables in BCS70 sweeps 1-10
- 18037 individuals in BCS70
- Exclude:
  - Routed variables.
  - Binary variables with prevalence <1%.
  - Variables with item non-response > 40%.
- Use summary scores for scales
- For non-response at sweep *t* we used the same 3 stage approach as in NCDS (using a bit stricter criteria)

Identifying predictors of non-response in NCDS

- For non-response at sweep t:
  - Stage 1: Univariable regressions for predictors at sweep 0, …, sweep $t - 1$. Retain predictors with $p < 0.001$.
  - Stage 2: Multivariable regressions for predictors at sweep 0, …, sweep $t - 1$ (separately). Retain predictors with $p < 0.05$.
  - Stage 3: MI; multivariable regressions for predictors at sweep 0, …, sweep $t - 1$, adjusted for predictors at previous waves. Retain predictors with $p < 0.001$

Variables sex, country of birth, participation in all previous sweeps and father's socioeconomic status are contained in the final set of variables and hence are not included in this formula

Flowchart presenting the number of variables used to find the predictors of non-response



Number of available variables: N=21021

Excluded "routed" variables (questions that depend on a specific response to a previous question), variables with >40% missing, variables with categories including >98% of cases, recoded variables with categories <1%, excluded all binary variables with prevalence <1% of cases if possible, and used index/score variables that combined information from many variables rather the individual constituent items

Variables to be used for Stage 1-3: N=967

Application of the Stage 1-3 process

Variables to be used after Stage 3:

Vary from N=7 (at sweep 2) to

N=16 (at sweep 10)

Predictors of non-response

	NR sweep 2 (age 5)	NR sweep 3 (age 10)	NR sweep 4 (age 16)	NR sweep 5 (age 26)	NR sweep 6 (age 30)	NR sweep 7 (age 34)	NR sweep 8 (age 38)	NR sweep 9 (age 42)	NR sweep 10 (age 46)
Sweep 1 (birth)	S1: 28 **S2: 15**	S1: 26 **S2: 14**	S1: 30 **S2: 13**	S1: 37 **S2: 15**	S1: 30 **S2: 14**	S1: 34 **S2: 13**	S1: 35 **S2: 13**	S1: 31 **S2: 12**	S1: 35 **S2: 11**
Sweep 2 (age 5)		S1: 15 **S2: 8**	S1: 39 **S2: 9**	S1: 56 **S2: 6**	S1: 49 **S2: 9**	S1: 55 **S2: 11**	S1: 57 **S2: 7**	S1: 49 **S2: 8**	S1: 51 **S2: 9**
Sweep 3 (age 10)			S1: 64 **S2: 15**	S1: 88 **S2: 14**	S1: 84 **S2: 12**	S1: 91 **S2: 13**	S1: 90 **S2: 13**	S1: 82 **S2: 9**	S1: 78 **S2: 9**
Sweep 4 (age 16)				S1: 63 **S2: 11**	S1: 39 **S2: 7**	S1: 57 **S2: 9**	S1: 65 **S2: 12**	S1: 49 **S2: 9**	S1: 54 **S2: 7**
Sweep 5 (age 26)					S1: 35 **S2: 7**	S1: 54 **S2: 8**	S1: 64 **S2: 13**	S1: 49 **S2: 5**	S1: 53 **S2: 10**
Sweep 6 (age 30)						S1: 56 **S2: 11**	S1: 75 **S2: 11**	S1: 56 **S2: 12**	S1: 64 **S2: 15**
Sweep 7 (age 34)							S1: 80 **S2: 10**	S1: 69 **S2: 17**	S1: 77 **S2: 16**
Sweep 8 (age 38)								S1: 23 **S2: 10**	S1: 27 **S2: 8**
Sweep 9 (age 42)									S1: 39 **S2: 5**
STAGE 3 without extra variables*	S3: 5 **(out of 15 in total from S2)**	S3: 7 **(out of 22 in total from S2)**	S3: 11 **(out of 37 in total from S2)**	S3: 10 **(out of 46 in total from S2)**	S3: 9 **(out of 49 in total from S2)**	S3: 12 **(out of 65 in total from S2)**	S3: 12 **(out of 79 in total from S2)**	S3: 9 **(out of 82 in total from S2)**	S3: 13 **(out of 90 in total from S2)**
Total with extra variables†	7	9	14	13	12	15	15	14	16

*We do not include in these counts the variables sex, country of birth, participation in all previous sweeps and father's socioeconomic status which were used directly in Stage 3

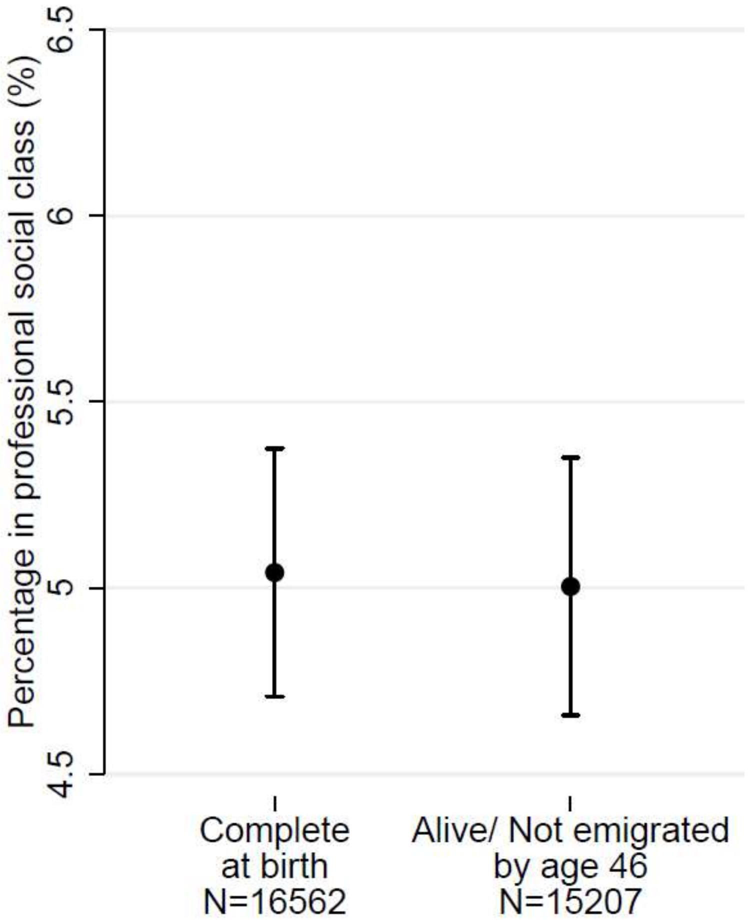
†We include in these counts the variables sex, country of birth, participation in all previous sweeps and father's socioeconomic status

Consistent predictors of participation

- Method of contraception (early sweeps)
- Paternal completion of education (early sweeps)
- Higher early life cognitive ability (sweeps in adulthood)
- Being female (sweeps in adulthood only)
- Few household moves (sweeps in adulthood only)
- Social participation – voting (sweeps in adulthood only)
- Home ownership. (sweeps in adulthood only)
- Parity, i.e. number of older siblings
- Participation in previous sweeps
- Paternal social class

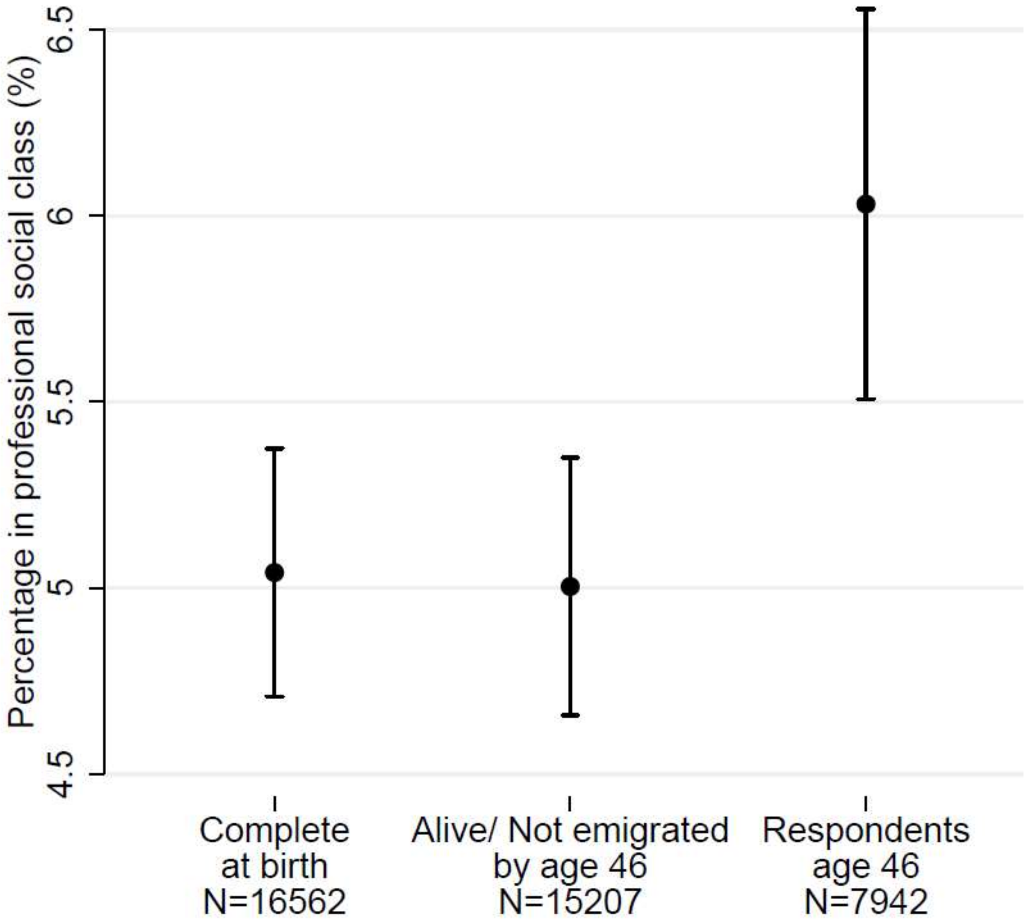
Internal validation: Paternal Social class - at birth

CENTRE FOR
LONGITUDINAL
STUDIES



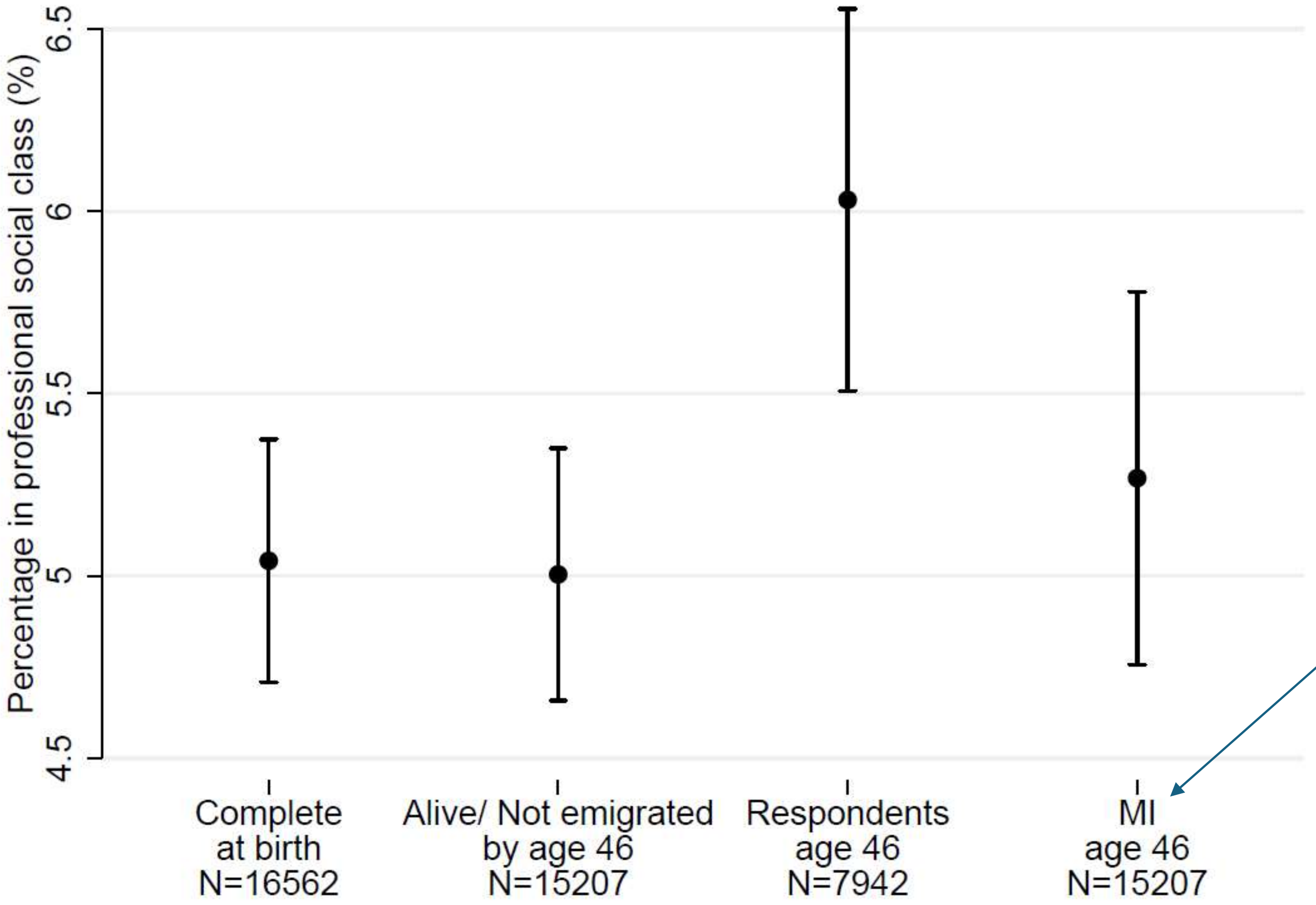
Internal validation: Paternal Social class - at birth

CENTRE FOR LONGITUDINAL STUDIES



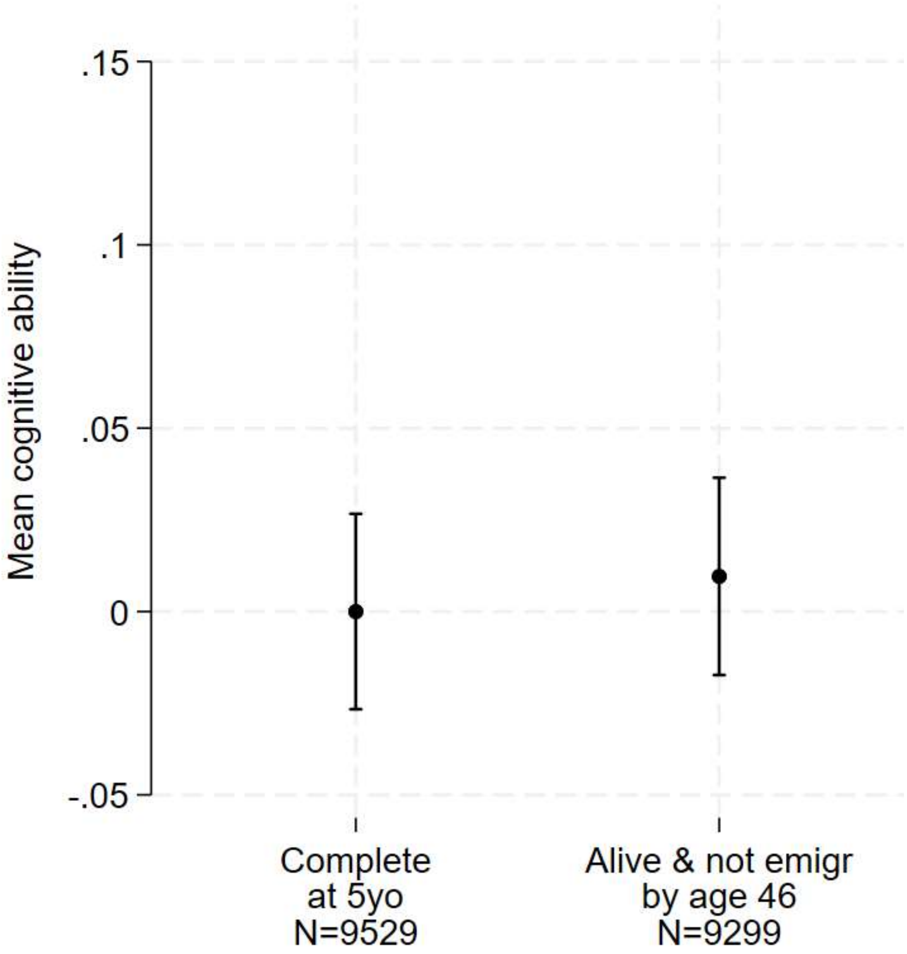
Internal validation: Paternal Social class - at birth

CENTRE FOR LONGITUDINAL STUDIES



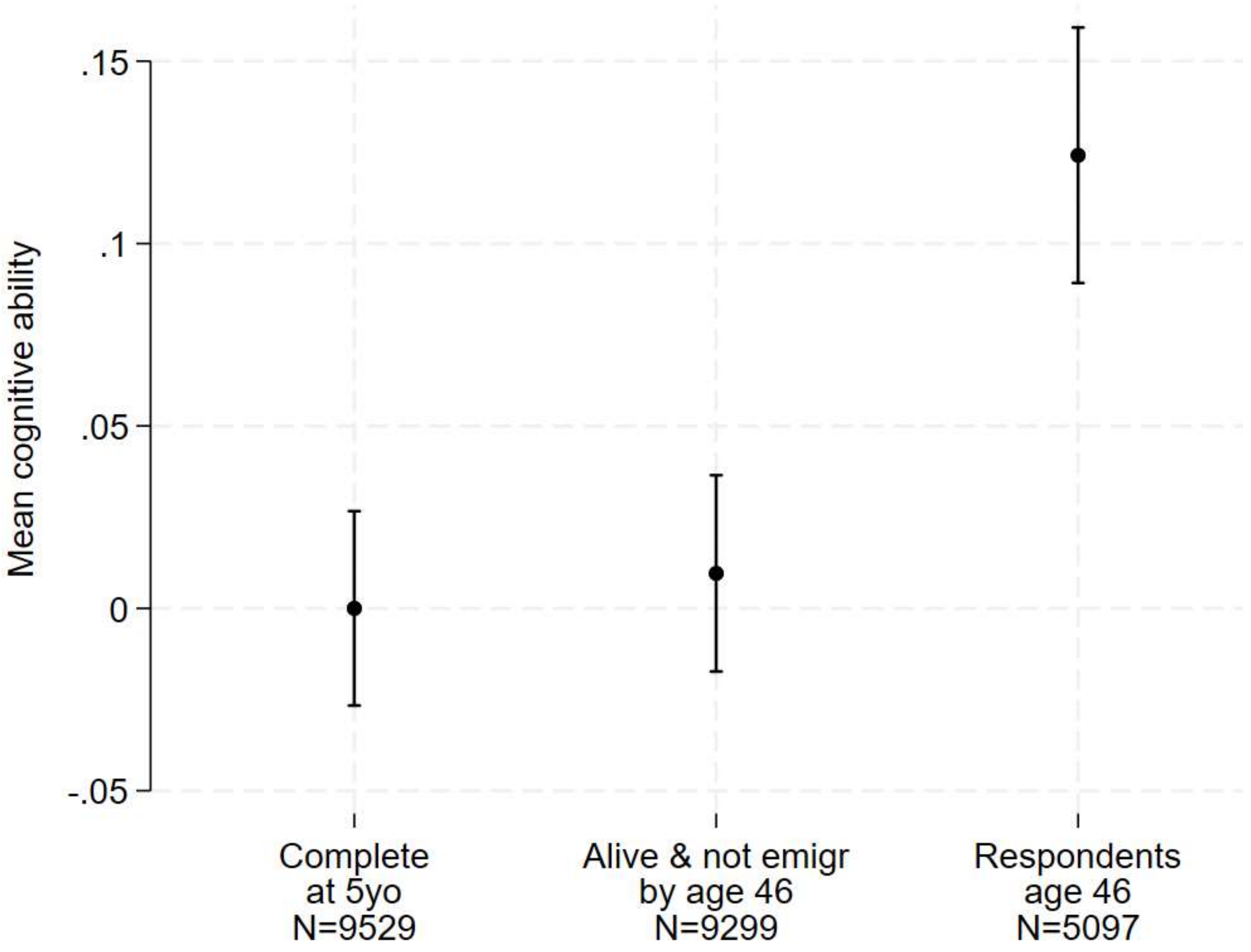
- Chained equations.
- 50 imputations.
- Auxiliary variables: All predictors of non-response at sweep 10 (age 46) from sweeps 0-9.

Internal validation: Cognitive ability - at 5yo



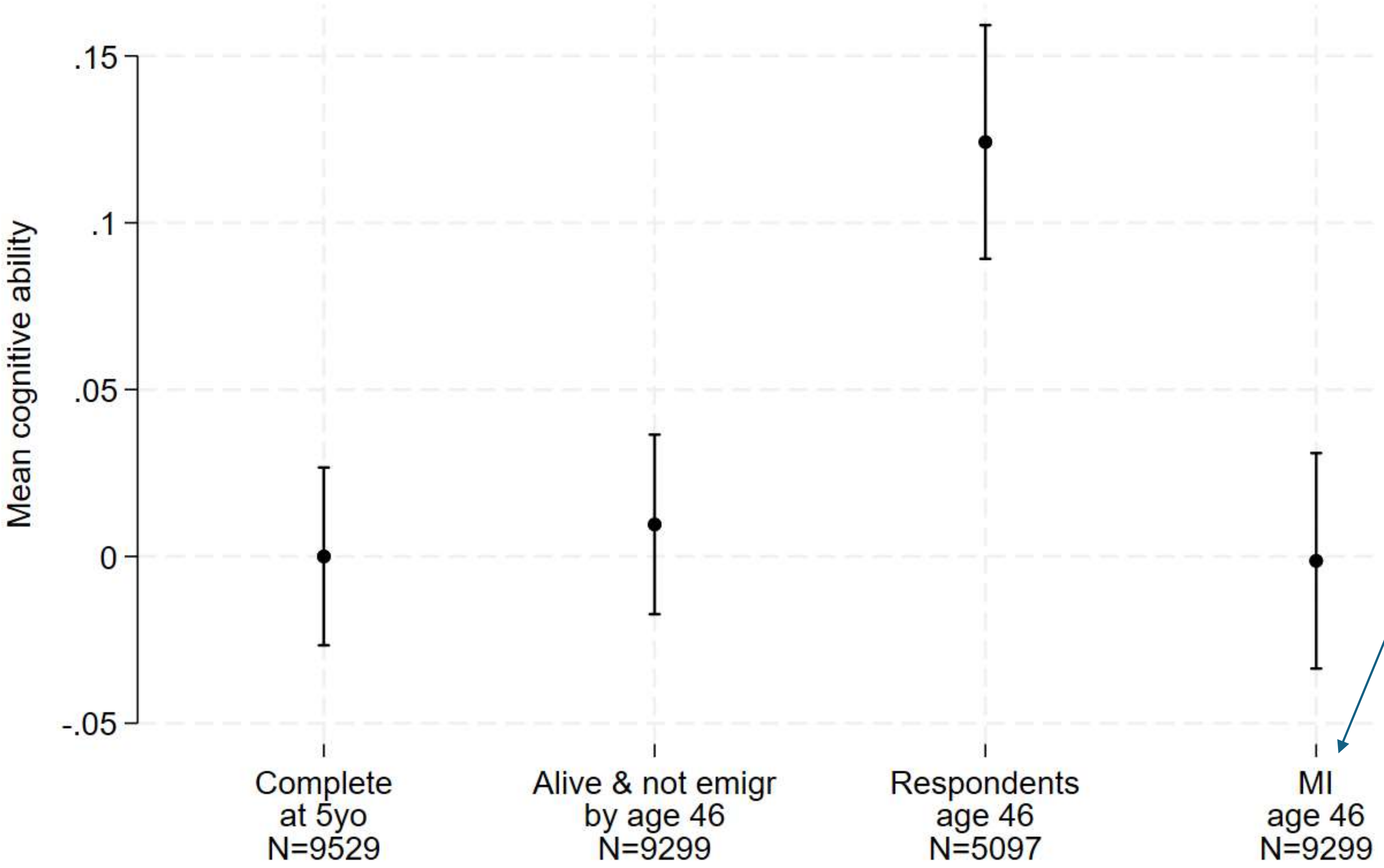
Internal validation: Cognitive ability - at 5yo

CENTRE FOR
LONGITUDINAL
STUDIES



Internal validation: Cognitive ability - at 5yo

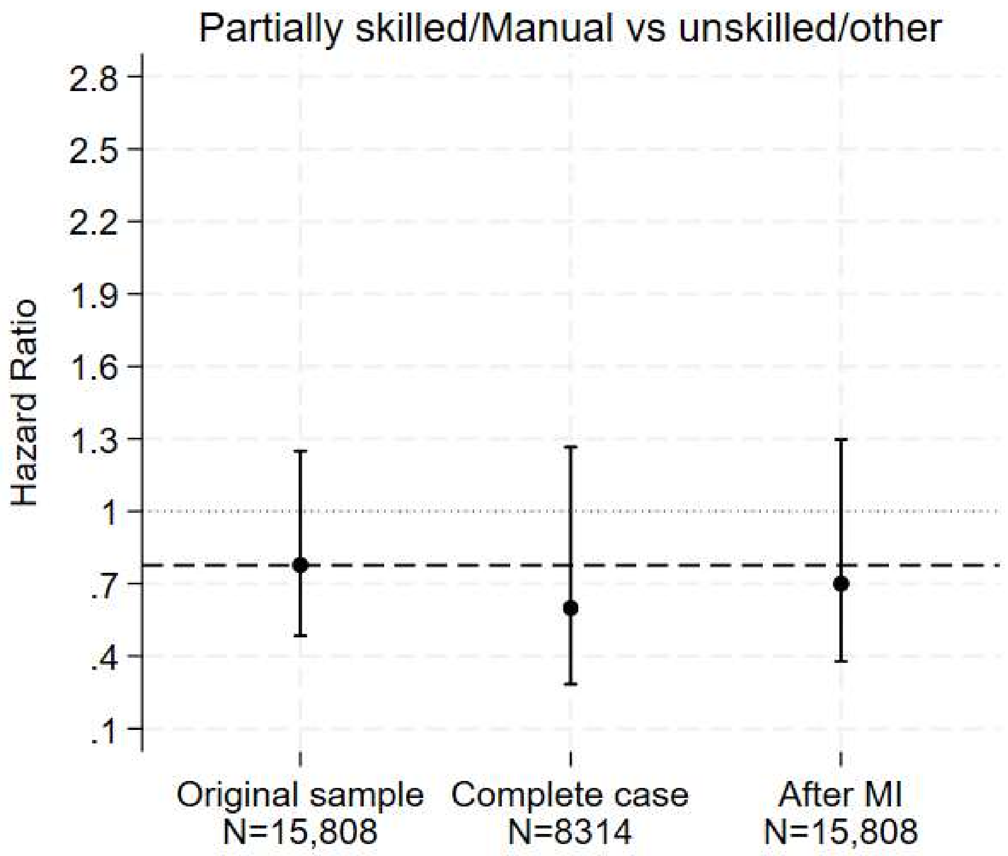
CENTRE FOR LONGITUDINAL STUDIES



- Chained equations.
- 50 imputations.
- Auxiliary variables: All predictors of non-response at sweep 10 (age 46) from sweeps 1-9.

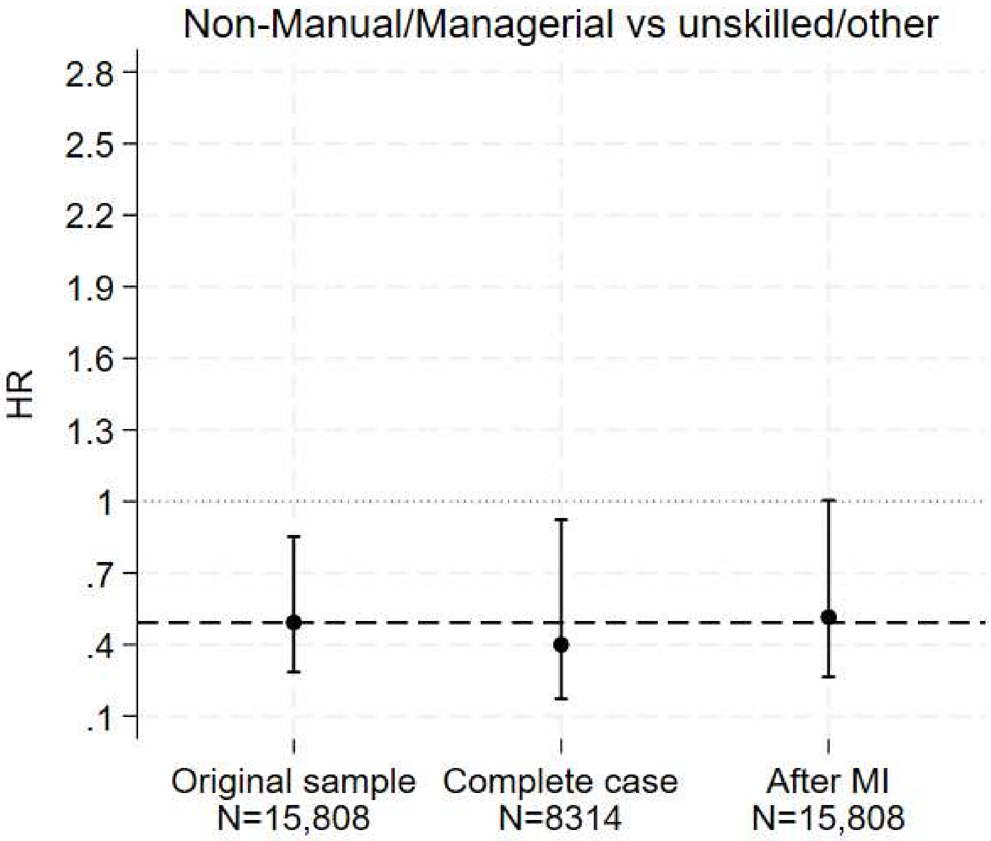
Hazard ratios of father's social class on participants' mortality after age 26 i) in the original sample ii) among 8314 respondents (complete case) of sweep 5 (age 26) and iii) after MI

CENTRE FOR
LONGITUDINAL
STUDIES

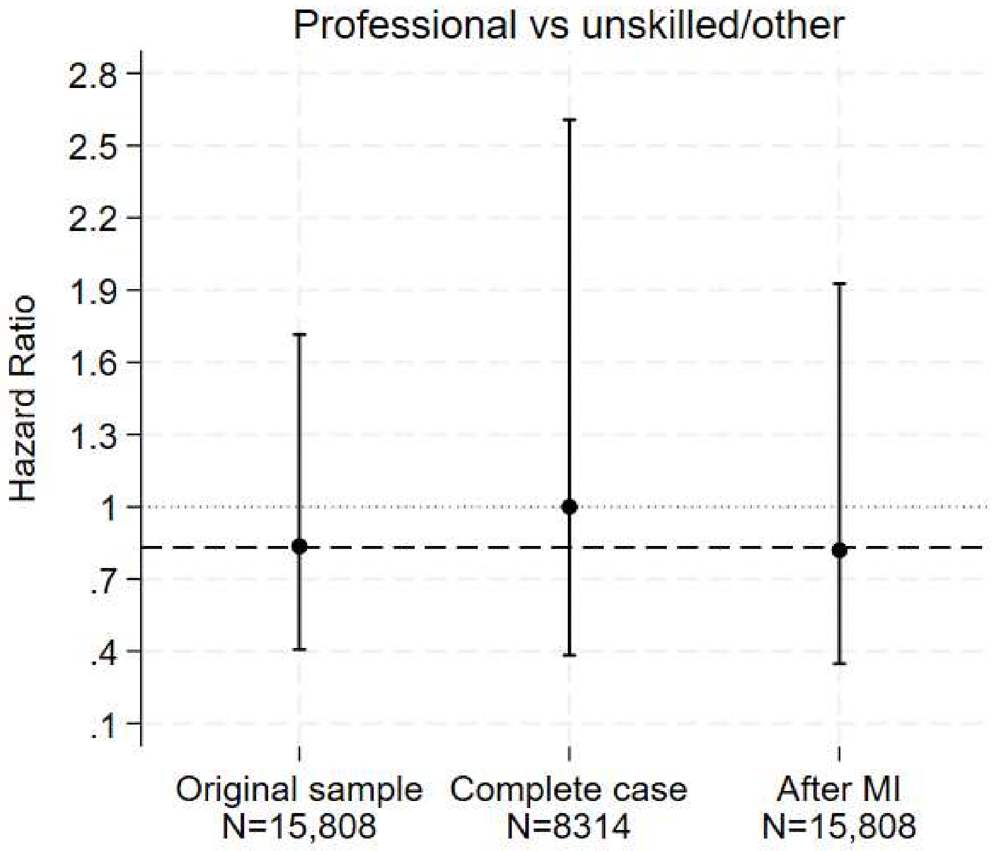


Hazard ratios of father's social class on participants' mortality after age 26 i) in the original sample ii) among 8314 respondents (complete case) of sweep 5 (age 26) and iii) after MI

CENTRE FOR LONGITUDINAL STUDIES



Hazard ratios of father’s social class on participants’ mortality after age 26 i) in the original sample ii) among 8314 respondents (complete case) of sweep 5 (age 26) and iii) after MI



Tackling missing data in BCS70

IN PRACTICE

CENTRE FOR LONGITUDINAL STUDIES

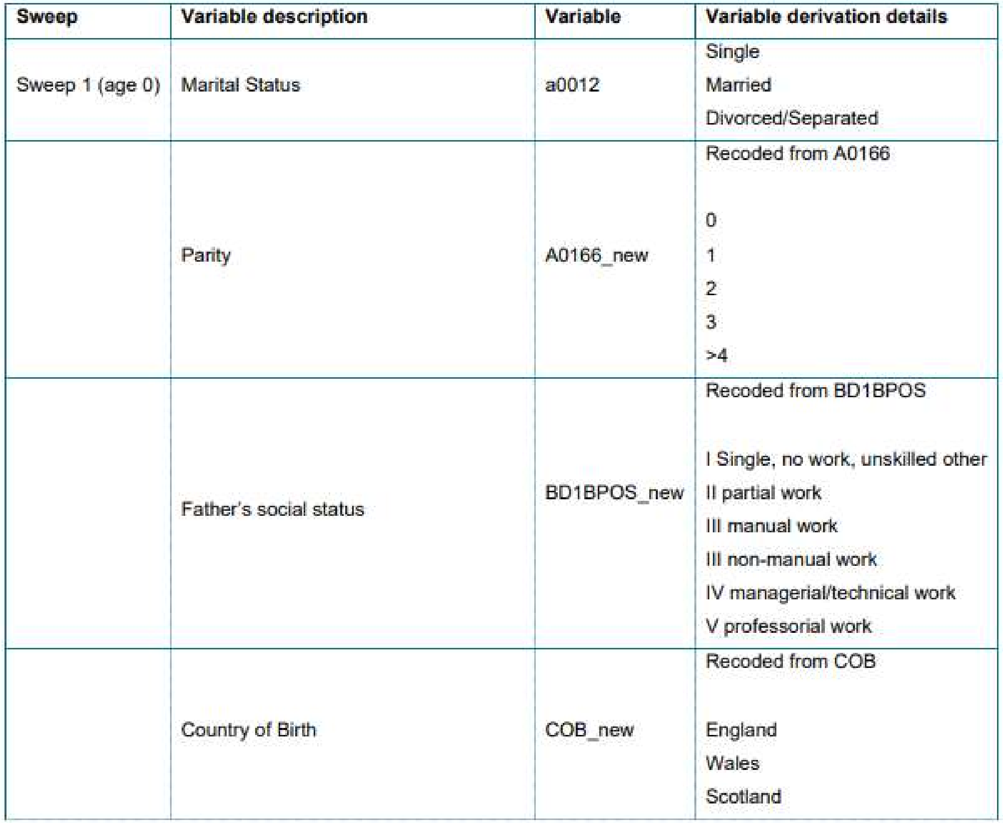
To address problems due to missing data in BCS70

- Utilise information from https://cls.ucl.ac.uk/wp-content/uploads/2020/04/Handling-Missing-Data-User-Guide-2024.pdf, pages 88-110
- Based on the sweep you want to tackle missing data, select the predictors of non-response you need
- E.g. when using MI, these variables can be used as auxiliary variables in your imputation model

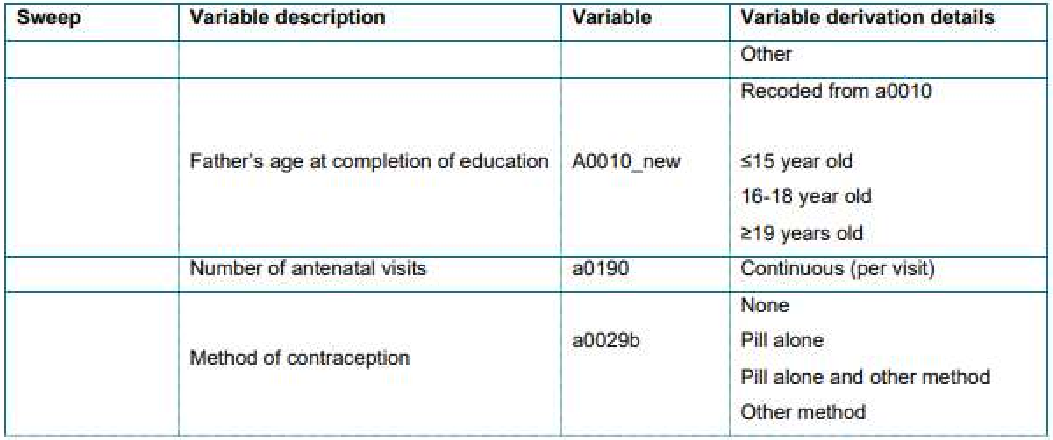
Tackling missing data in BCS70

IN PRACTICE – sweep 2 (age 5)

Table B1. Predictors of non-response at Sweep 2 (age 5).



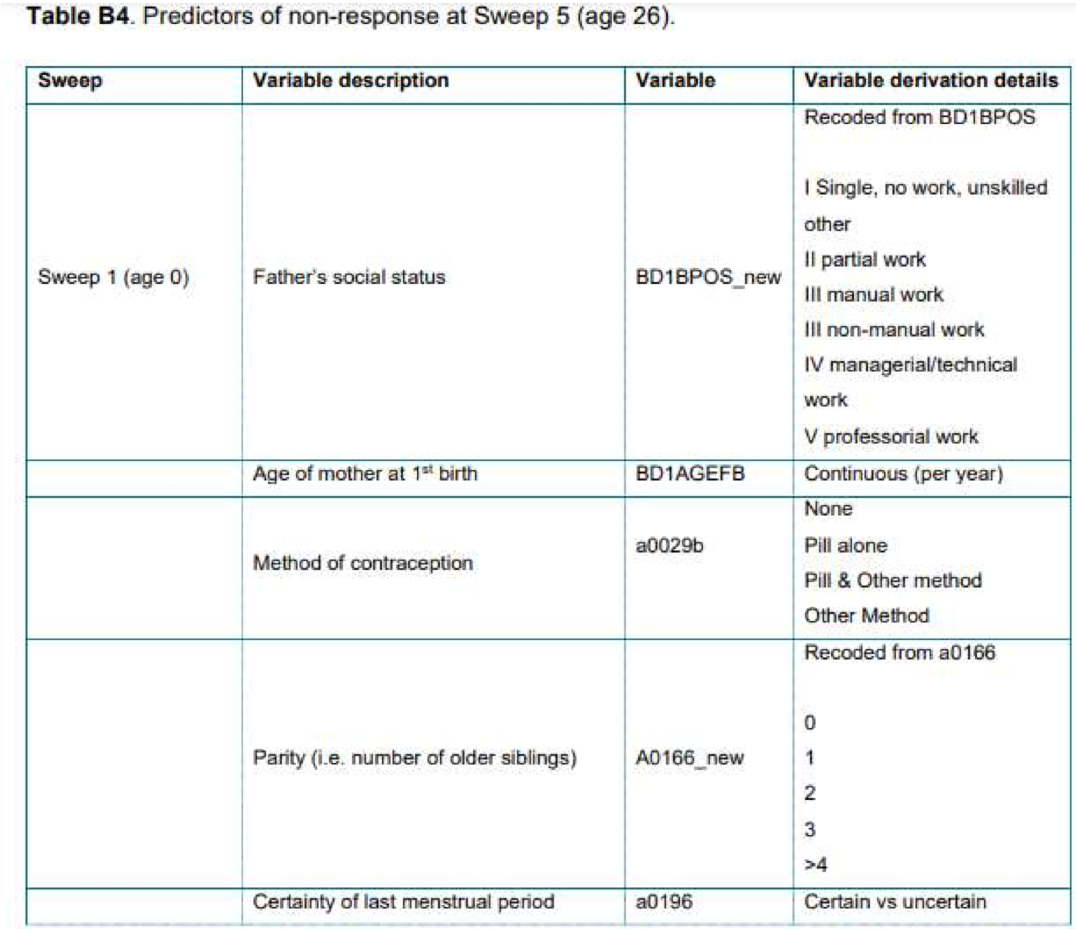
Sweep	Variable description	Variable	Variable derivation details
Sweep 1 (age 0)	Marital Status	a0012	Single Married Divorced/Separated
	Parity	A0166_new	Recoded from A0166 0 1 2 3 >4
	Father's social status	BD1BPOS_new	Recoded from BD1BPOS I Single, no work, unskilled other II partial work III manual work III non-manual work IV managerial/technical work V professorial work
	Country of Birth	COB_new	Recoded from COB England Wales Scotland



Sweep	Variable description	Variable	Variable derivation details
			Other
	Father's age at completion of education	A0010_new	Recoded from a0010 ≤15 year old 16-18 year old ≥19 years old
	Number of antenatal visits	a0190	Continuous (per visit)
	Method of contraception	a0029b	None Pill alone Pill alone and other method Other method

Tackling missing data in BCS70

IN PRACTICE – sweep 5 (age 26)



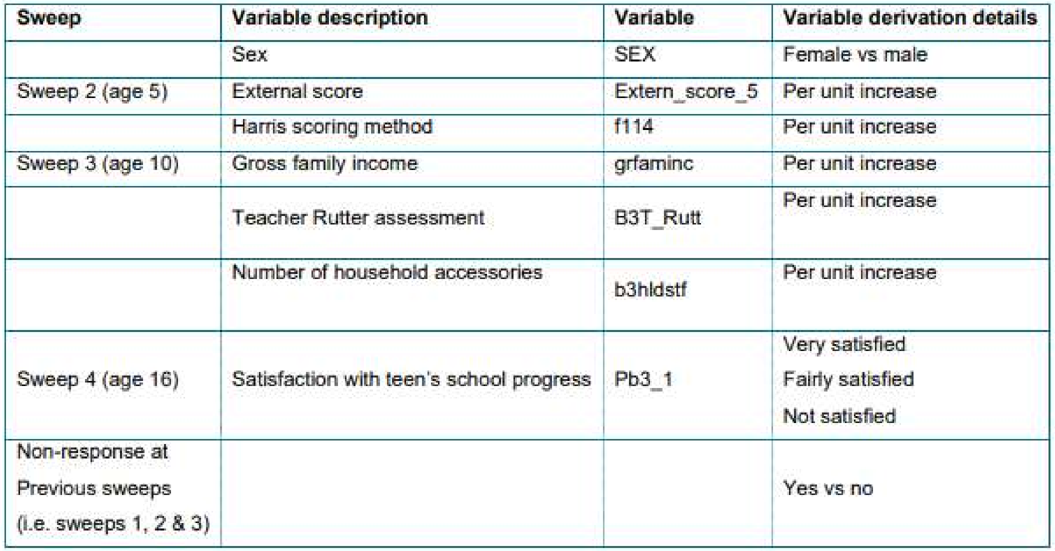


Table B4. Predictors of non-response at Sweep 5 (age 26).

Sweep	Variable description	Variable	Variable derivation details
Sweep 1 (age 0)	Father's social status	BD1BPOS_new	Recoded from BD1BPOS I Single, no work, unskilled other II partial work III manual work III non-manual work IV managerial/technical work V professorial work
	Age of mother at 1st birth	BD1AGEFB	Continuous (per year)
	Method of contraception	a0029b	None Pill alone Pill & Other method Other Method
	Parity (i.e. number of older siblings)	A0166_new	Recoded from a0166 0 1 2 3 >4
	Certainty of last menstrual period	a0196	Certain vs uncertain
	Sex	SEX	Female vs male
Sweep 2 (age 5)	External score	Extern_score_5	Per unit increase
	Harris scoring method	f114	Per unit increase
Sweep 3 (age 10)	Gross family income	grfaminc	Per unit increase
	Teacher Rutter assessment	B3T_Rutt	Per unit increase
	Number of household accessories	b3hldstf	Per unit increase
Sweep 4 (age 16)	Satisfaction with teen's school progress	Pb3_1	Very satisfied Fairly satisfied Not satisfied
Non-response at Previous sweeps (i.e. sweeps 1, 2 & 3)			Yes vs no

Tackling missing data in BCS70

IN PRACTICE – sweep 10 (age 46)

CENTRE FOR
LONGITUDINAL
STUDIES

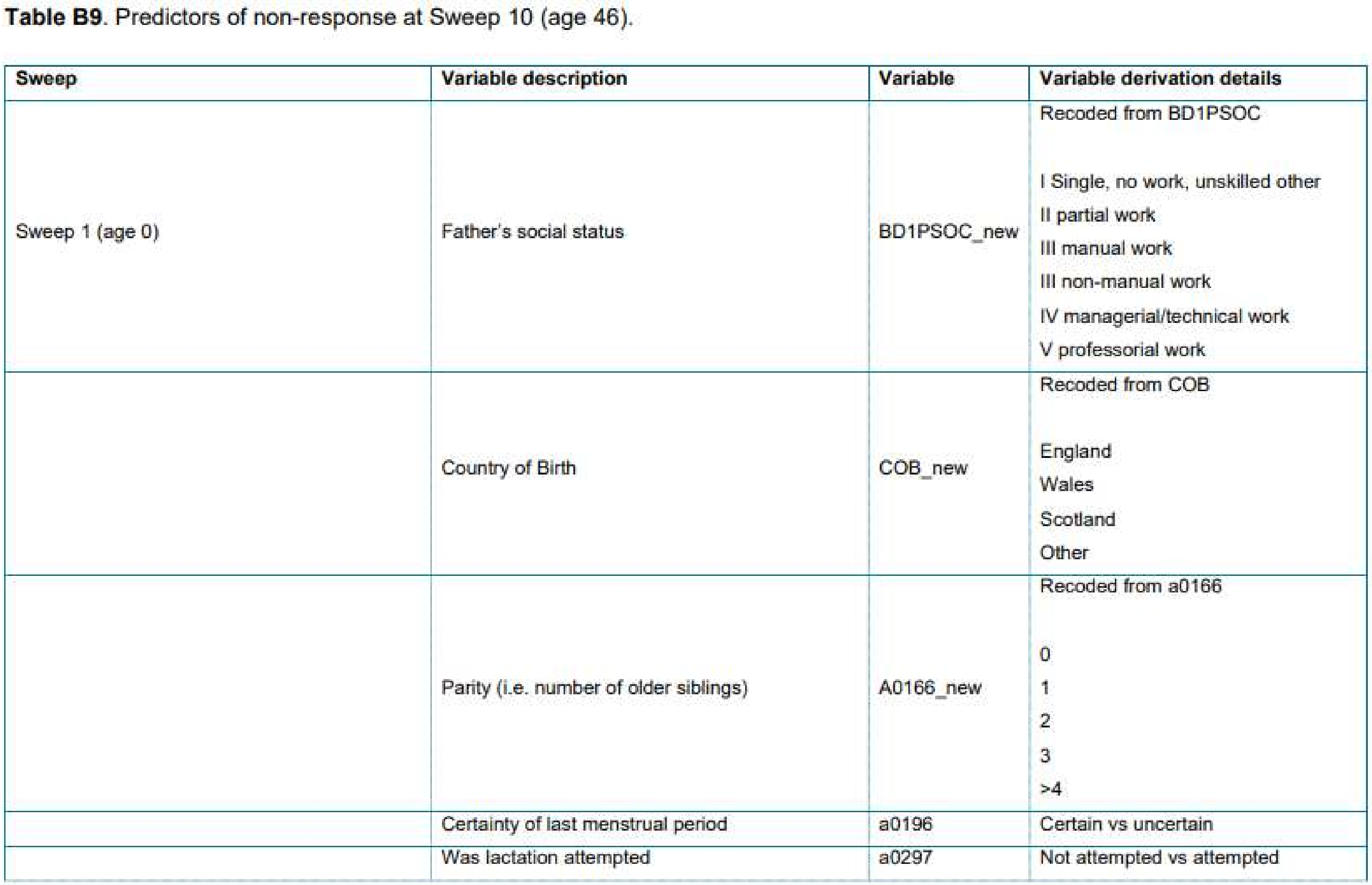
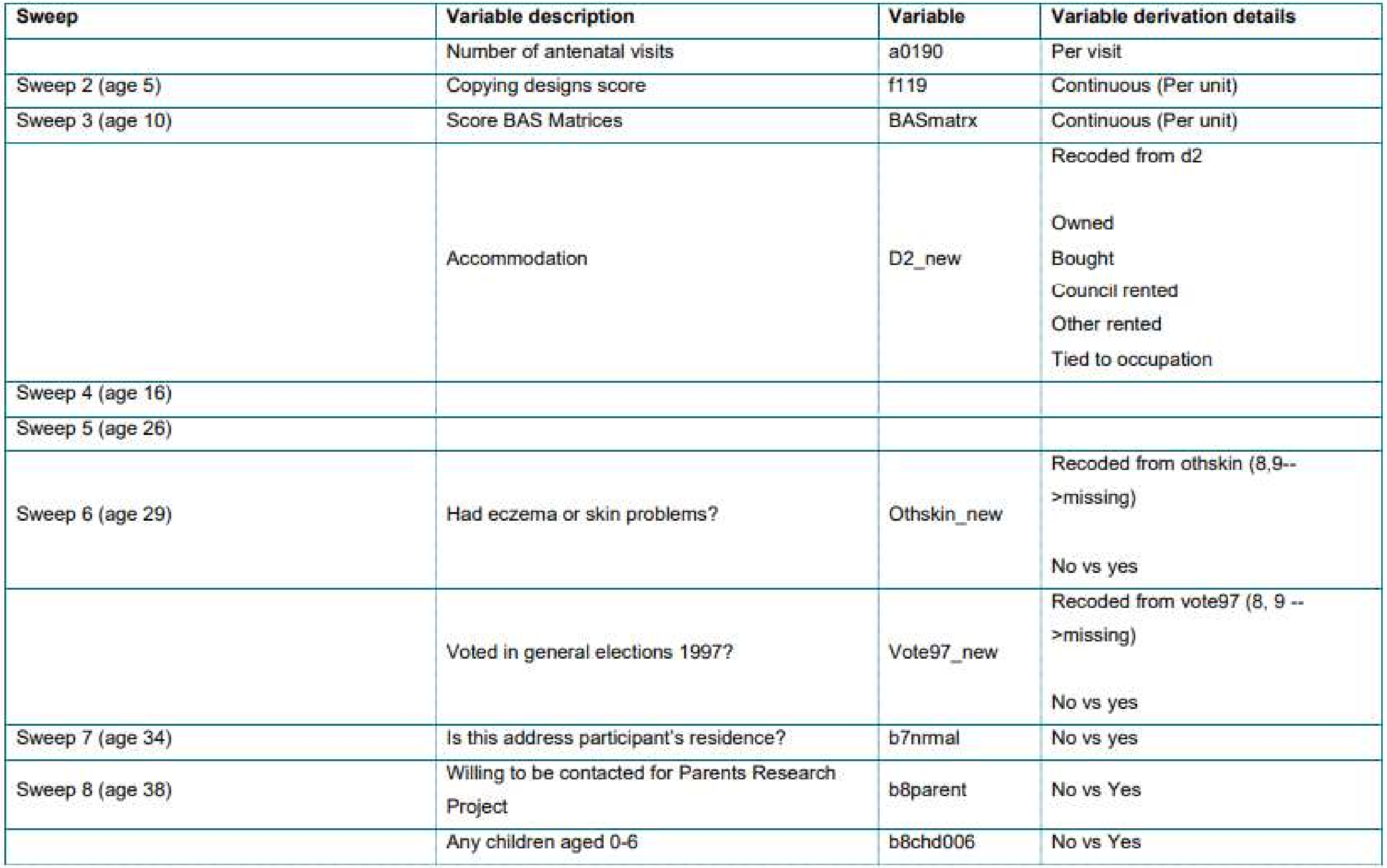


Table B9. Predictors of non-response at Sweep 10 (age 46).

Sweep	Variable description	Variable	Variable derivation details
Sweep 1 (age 0)	Father's social status	BD1PSOC_new	Recoded from BD1PSOC I Single, no work, unskilled other II partial work III manual work III non-manual work IV managerial/technical work V professorial work
	Country of Birth	COB_new	Recoded from COB England Wales Scotland Other
	Parity (i.e. number of older siblings)	A0166_new	Recoded from a0166 0 1 2 3 >4
	Certainty of last menstrual period	a0196	Certain vs uncertain
	Was lactation attempted	a0297	Not attempted vs attempted

Tackling missing data in BCS70

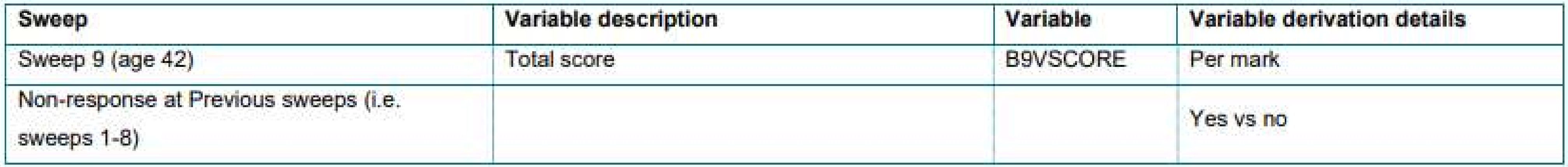
IN PRACTICE – sweep 10 (age 46) - continued



Sweep	Variable description	Variable	Variable derivation details
	Number of antenatal visits	a0190	Per visit
Sweep 2 (age 5)	Copying designs score	f119	Continuous (Per unit)
Sweep 3 (age 10)	Score BAS Matrices	BASmatrx	Continuous (Per unit)
	Accommodation	D2_new	Recoded from d2 Owned Bought Council rented Other rented Tied to occupation
Sweep 4 (age 16)			
Sweep 5 (age 26)			
Sweep 6 (age 29)	Had eczema or skin problems?	Othskin_new	Recoded from othskin (8,9-->missing) No vs yes
	Voted in general elections 1997?	Vote97_new	Recoded from vote97 (8, 9 -->missing) No vs yes
Sweep 7 (age 34)	Is this address participant's residence?	b7nrmal	No vs yes
Sweep 8 (age 38)	Willing to be contacted for Parents Research Project	b8parent	No vs Yes
	Any children aged 0-6	b8chd006	No vs Yes

Tackling missing data in BCS70

IN PRACTICE – sweep 10 (age 46) - continued



Sweep	Variable description	Variable	Variable derivation details
Sweep 9 (age 42)	Total score	B9VSCORE	Per mark
Non-response at Previous sweeps (i.e. sweeps 1-8)			Yes vs no

Tackling missing data in BCS70

IN PRACTICE

These variables can be used to restore representativeness and

- Present summary statistics of the characteristics of individuals
- Assess causal relationships between your exposure and your outcome, account for potential confounders

Summary

CENTRE FOR LONGITUDINAL STUDIES

- We have identified variables which predict non-response at each wave of BCS70.
- These can be used as auxiliary variables in subsequent analyses to increase the plausibility of the MAR assumption.
- Simple test analyses have shown this approach to perform well.
- Lists of predictors of non-response available via Handling missing data in the CLS cohort studies (https://cls.ucl.ac.uk/wp-content/uploads/2020/04/Handling-Missing-Data-User-Guide-2024.pdf).
- Will be updated when new waves of data become available.
- Our work will facilitate researchers who plan to utilise BCS70 and help them address bias due to missing data

CENTRE FOR
LONGITUDINAL
STUDIES

Q&A

Thank you

CENTRE FOR
LONGITUDINAL
STUDIES

Economic
and Social
Research Council