Using linked Hospital Episode Statistics data to aid the handling of missing cohort data

Nasir Rajah and Richard Silverwood

Health Studies User Conference

July 2022

CENTRE FOR LONGITUDINAL STUDIES

Economic and Social Research Council

Co-investigators

CENTRE FOR LONGITUDINAL STUDIES

Lisa Calderwood, George Ploubidis

Centre for Longitudinal Studies, UCL Social Research Institute

Bianca De Stavola, Katie Harron

Population, Policy & Practice Department, UCL Great Ormond Street Institute of Child Health

Outline

CENTRE FOR LONGITUDINAL STUDIES

1. Background
2. Data sources: NCDS, HES & linkage
3. HES predictors of NCDS non-response
4. Restoring NCDS sample representativeness (preliminary results)
5. Conclusions

CENTRE FOR LONGITUDINAL STUDIES

Background

Missing data

CENTRE FOR LONGITUDINAL STUDIES

- Non-response is common in longitudinal surveys.
- Missing values due to non-response mean less efficient estimates because of the reduced size of the of the analysis sample.
- Also introduce the potential for bias since respondents are often systematically different from non-respondents.

Analytical Strategy

CENTRE FOR LONGITUDINAL STUDIES

- Growing interest in whether linked administrative data have the potential to aid analyses subject to missing data in cohort studies.
- Identify predictors of cohort non-response in linked administrative data.
- Explore whether added value in including identified variables as auxiliary variables with respect to restoring sample representativeness.
- Today: NCDS and HES.

CENTRE FOR LONGITUDINAL STUDIES

Data sources: NCDS & HES

1958 National Child Development Study (NCDS)

CENTRE FOR LONGITUDINAL STUDIES

- Longitudinal birth cohort study of all babies born in a single week in Great Britain in 1958.
- Initial N = 17,415, later augmented by immigrants born in target week.
- Multidisciplinary content spanning physical and educational development, economic circumstances, employment, family life, health behaviour, wellbeing, social participation and attitudes.

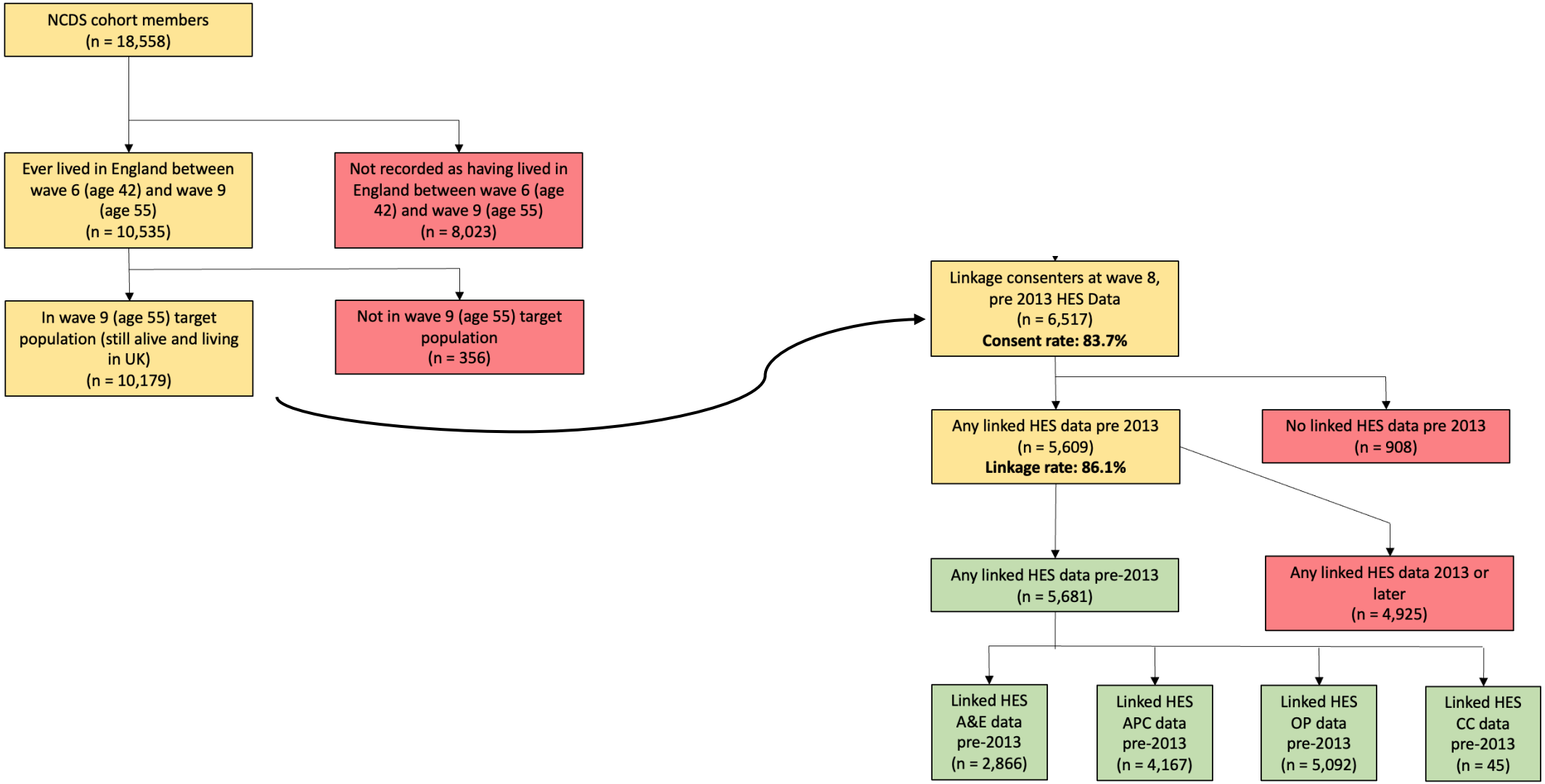
Hospital Episode Statistics (HES)

CENTRE FOR LONGITUDINAL STUDIES

- A collection of databases containing details of interactions with NHS hospitals in England:
  - Admitted Patient Care (APC)
  - Critical Care (CC) admissions
  - Accident and Emergency (A&E) attendances
  - Outpatient (OP) appointments
- Datasets include dates, diagnoses, procedures, patient demographics, and hospital characteristics for each hospital episode.
- Often multiple episodes per admission (APC).
- Period of data availability differs by dataset: APC (1997-), OP (2003-), A&E (2007-), CC (2009-).
- Linkage between NCDS and all four HES datasets undertaken on the basis of consent at sweep 8 (age 50).

NCDS-HES linkage

CENTRE FOR LONGITUDINAL STUDIES



CENTRE FOR LONGITUDINAL STUDIES

HES predictors of NCDS non-response

Potential predictors of non-response

CENTRE FOR LONGITUDINAL STUDIES

- Variables derived using HES APC, OP and A&E Prior to NCDS9 (2013).
- A total of 58 variables derived relating to:
  - Numbers of admissions and appointments
  - Missed appointments
  - Investigations undertaken
  - Diagnoses
  - Treatments received
- Assume cohort members who were eligible for and consented to linkage but did not have linked data truly did not have a relevant interaction with an NHS hospital.
- E.g. No linked HES APC data → truly no hospital admissions → all APC-based diagnoses and treatments = "No".

Identifying predictors of non-response

CENTRE FOR LONGITUDINAL STUDIES

- Least absolute shrinkage and selection operator (LASSO) on identified HES variables (58)
- LASSO removes variables that are not influential in predicting non-response at age 55.
- Uses a penalty (lambda) that is determined by cross-validation
- Select lambda value that gives the minimum mean cross validated error (minimum misclassification error)
- Variables selected after LASSO include:
  - Number of A&E Appointments (continuous)
  - Treatment for adult mental illness in APC (binary)
  - Proportion of appointments missed in outpatient (continuous)
  - 5 ICD Chapter Diagnoses in APC (e.g., ICD Chapter IV: Endocrine, nutritional and metabolic diseases) (binary)
  - 2 Operation Codes in APC (e.g., Operation Code T – Soft Tissue) (binary)

CENTRE FOR LONGITUDINAL STUDIES

Restoring NCDS sample representativeness

CENTRE FOR LONGITUDINAL STUDIES

We include the ten HES variables identified as being predictive of non-response as auxiliary variables in MI analyses.

We explore the performance of this approach in two separate samples

A. Amongst HES Linkage Consenters

B. Amongst the Whole Cohort

CENTRE FOR LONGITUDINAL STUDIES

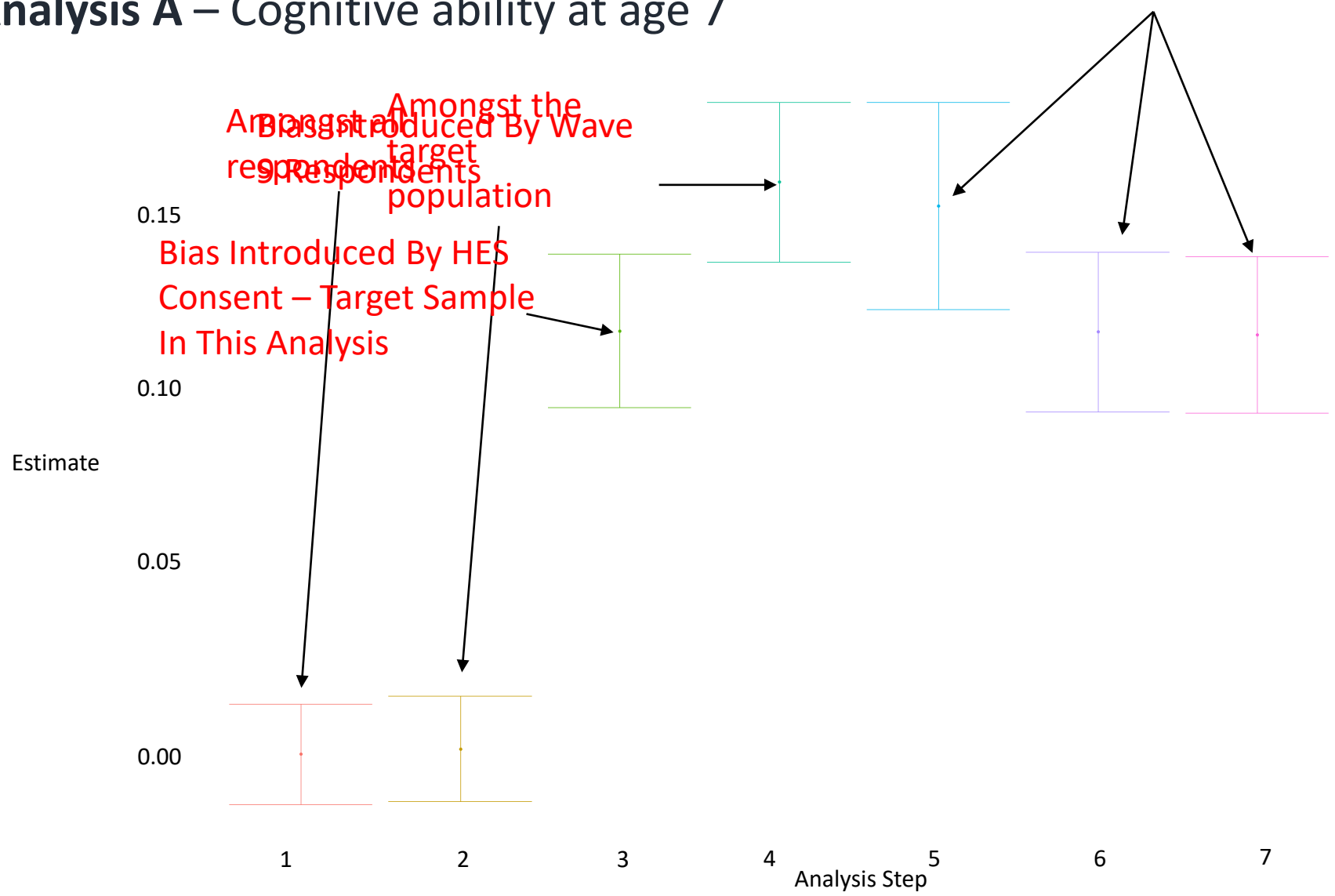
Analysis A – HES Consenters

1. See whether distributions of variables from earlier sweeps can be replicated using only data from respondents at a later sweep

Restoring NCDS sample representativeness

CENTRE FOR LONGITUDINAL STUDIES

Analysis A – Cognitive ability at age 7



1 – All Respondents
n = 14,407

2 – Target Population
n = 12,938

3 – Consenters
n = 5,546

4 – Wave 9 Respondents
n = 4,928

5 – MI HES
n = 5,546

6 – MI Survey
n = 5,546

7 – MI Survey & HES
n = 5,546

CENTRE FOR LONGITUDINAL STUDIES

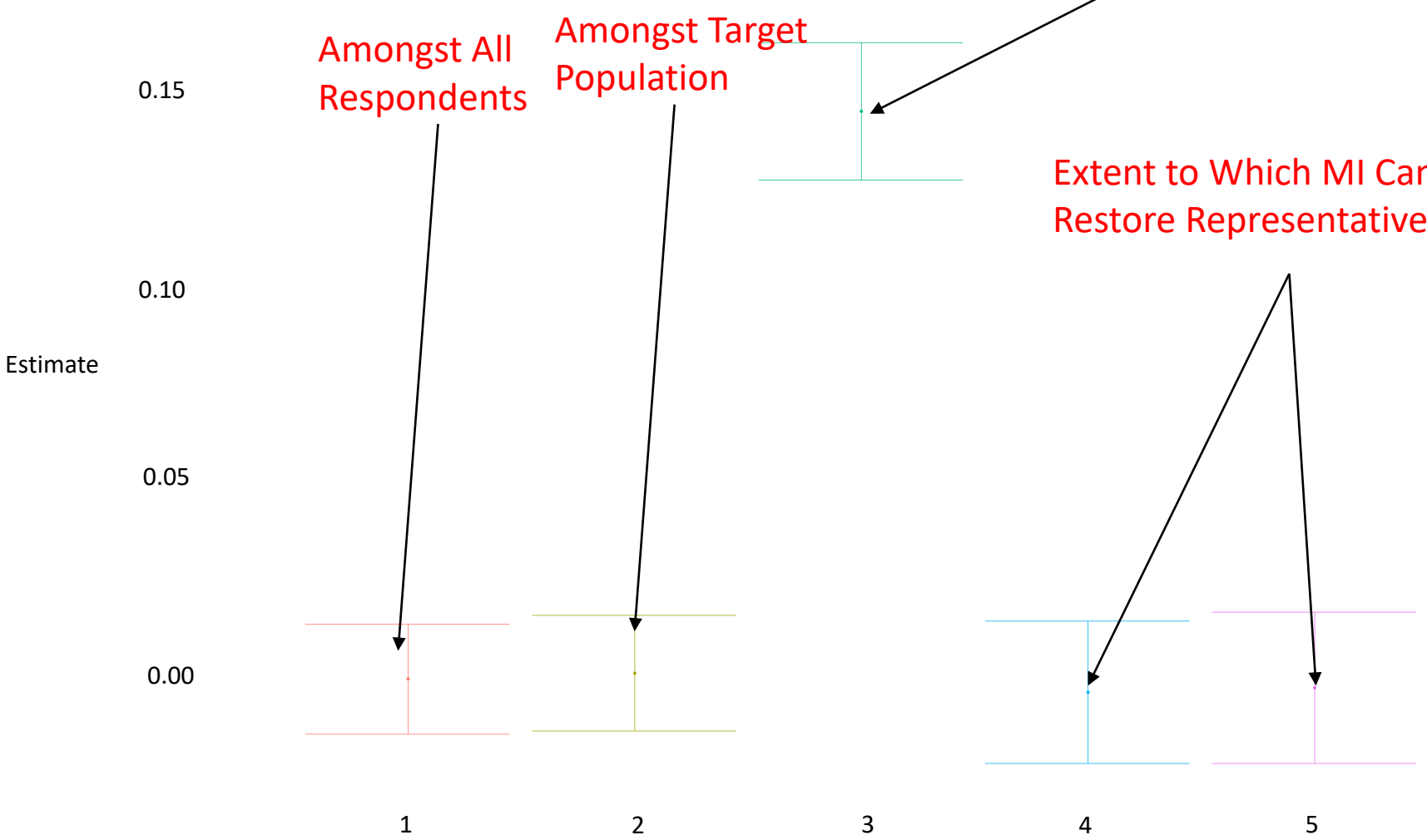
Analysis B – Amongst the Whole Cohort

1. See whether distributions of variables from earlier sweeps can be replicated using only data from respondents at a later sweep.

Restoring NCDS sample representativeness

CENTRE FOR LONGITUDINAL STUDIES

Analysis B – Cognitive ability at age 7



1 – All Respondents
n = 14,407

2 – Target Population
n = 12,938

3 – Wave 9 Respondents
n = 7,839

4 – MI Survey
n = 12,938

5 – MI Survey & HES
n = 12,938

CENTRE FOR
LONGITUDINAL
STUDIES

Conclusions

Conclusions

CENTRE FOR LONGITUDINAL STUDIES

- We have identified HES variables which are predictive of non-response at NCDS wave 9 (age 55).
- Incorporating these variables as auxiliary variables in MI analyses of NCDS had relatively limited impact on restoring sample representativeness.
- We found no additional gain relative to using only previously identified survey predictors of non-response.
- Whilst this finding may not extend to other analyses or NCDS sweeps, it highlights the utility of survey variables in handling non-response.
- This provides a straightforward approach for missing data handling, which is easily implemented in standard software.

References

CENTRE FOR LONGITUDINAL STUDIES

- 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.
- 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. J Clin Epidemiol. 2021;136:44-54.
- 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.
- Silverwood, R., Rajah, N., Calderwood, L., De Stavola, B.L., Harron, K., Ploubidis, G.B. (2022) Examining the quality and sample representativeness of linked survey and administrative data: linking the 1958 National Child Development Study to Hospital Episode Statistics data. CLS Working Paper 2022/5. London: UCL Centre for Longitudinal Studies.
- 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. 2021.

Funding

CENTRE FOR LONGITUDINAL STUDIES

- This work was funded by the Economic & Social Research Council (ES/V006037/1).
- The Centre for Longitudinal Studies is supported by the Economic & Social Research Council (ES/M001660/1).

Thank you.

CENTRE FOR LONGITUDINAL STUDIES

Economic and Social Research Council