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
Co-investigators

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

• Missing values due to non-response mean less efficient estimates because of the reduced size of the analysis sample.

• Also introduce the potential for bias since respondents are often systematically different from non-respondents.
Analytical Strategy

- 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.
Data sources: NCDS & HES
1958 National Child Development Study (NCDS)

- 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)

- 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).
- Linkage between NCDS and all four HES datasets undertaken on the basis of consent at sweep 8 (age 50).
NCDS-HES linkage

NCDS cohort members
(n = 18,558)

- Ever lived in England between wave 6 (age 42) and wave 9 (age 55)
  (n = 10,535)
- In wave 9 (age 55) target population (still alive and living in UK)
  (n = 10,179)
- Not recorded as having lived in England between wave 6 (age 42) and wave 9 (age 55)
  (n = 8,023)
- Not in wave 9 (age 55) target population
  (n = 356)

Linkage consenters at wave 8, pre 2013 HES Data
(n = 6,517)
Consent rate: 83.7%

- Any linked HES data pre 2013
  (n = 5,609)
  Linkage rate: 86.1%
- No linked HES data pre 2013
  (n = 908)

- Any linked HES data pre-2013
  (n = 5,681)
- Any linked HES data 2013 or later
  (n = 4,925)

- Linked HES A&E data pre-2013
  (n = 2,866)
- Linked HES APC data pre-2013
  (n = 4,167)
- Linked HES OP data pre-2013
  (n = 5,092)
- Linked HES CC data pre-2013
  (n = 45)
HES predictors of NCDS non-response
Potential predictors of non-response

• 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

- 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)
Restoring NCDS sample representativeness
Restoring NCDS sample representativeness

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
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

Analysis A – Cognitive ability at age 7

Estimate

Analysis Step

1 2 3 4 5 6 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

Extent to Which MI Can Restore Representativeness
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

**Analysis B – Cognitive ability at age 7**

Amongst All Respondents

Amongst Target Population

Bias Introduced By Restricting to Wave 9 Respondents

<table>
<thead>
<tr>
<th>Analysis Step</th>
<th>Description</th>
<th>Sample Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>All Respondents</td>
<td>n = 14,407</td>
</tr>
<tr>
<td>2</td>
<td>Target Population</td>
<td>n = 12,938</td>
</tr>
<tr>
<td>3</td>
<td>Wave 9 Respondents</td>
<td>n = 7,839</td>
</tr>
<tr>
<td>4</td>
<td>MI Survey</td>
<td>n = 12,938</td>
</tr>
<tr>
<td>5</td>
<td>MI Survey &amp; HES</td>
<td>n = 12,938</td>
</tr>
</tbody>
</table>

Extent to Which MI Can Restore Representativeness
Conclusions
Conclusions

- 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.


Funding

• 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.