CLS Missing Data Strategy

Richard Silverwood & Michalis Katsoulis

27 April 2023
Outline

1. CLS Missing Data Strategy
2. 1958 National Child Development Study (NCDS)
3. 1970 British Cohort Study (BCS70)
4. Other CLS studies
5. COVID-19 Surveys
6. Leveraging data linkages
7. Resources
CLS Missing Data Strategy
CLS Missing Data Strategy

• All users of longitudinal data need to consider the issue of missing data since some non-response is inevitable.
• Strategies for how to deal with missing data depend on the nature of non-response.
• 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), implying that systematic differences between the missing values and the observed values can be explained by observed data.
CLS Missing Data Strategy

• Most studies 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 sweep in each study.

• This allows us to capitalise on the rich data cohort members have provided over the years/decades in order to deal with missing data and reduce bias.
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 (NCDS)
Response in NCDS
Identifying predictors of non-response in NCDS

• Aim to maximise the plausibility of the MAR assumption by exploiting the richness of NCDS data.
• Using a data driven approach we identify the variables that are associated with non-response at each sweep.
• These can then be used as auxiliary variables.
• Substantive interest in understanding the drivers of non-response within and between cohorts.
Identifying predictors of non-response in NCDS

• ~17,500 variables in NCDS sweeps 0-8.
• Exclude:
  • Routed variables.
  • Binary variables with prevalence <1%.
  • Variables with item non-response > 50%.
• Use summary scores for scales.
• Use summary measures; exclude constituent variables.
• 587 variables meeting inclusion criteria
  → multi-stage, data driven approach.
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 $p < 0.05$.
  • Stage 2: Multivariable regressions for predictors at sweep 0, …, sweep $t-1$. 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 $p < 0.001$.

• Full details in Mostafa et al (2021) and NCDS Missing Data User Guide.
## Predictors of non-response

<table>
<thead>
<tr>
<th></th>
<th>NR sweep 1 (age 7)</th>
<th>NR sweep 2 (age 11)</th>
<th>NR sweep 3 (age 16)</th>
<th>NR sweep 4 (age 23)</th>
<th>NR sweep 5 (age 33)</th>
<th>NR sweep 6 (age 42)</th>
<th>NR BM sweep (age 44)</th>
<th>NR sweep 7 (age 46)</th>
<th>NR sweep 8 (age 50)</th>
<th>NR sweep 9 (age 55)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sweep 0 (birth)</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>5</td>
<td>3</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>Sweep 1 (age 7)</td>
<td>5</td>
<td>3</td>
<td>3</td>
<td>5</td>
<td>1</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Sweep 2 (age 11)</td>
<td></td>
<td>1</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Sweep 3 (age 16)</td>
<td></td>
<td></td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>Sweep 4 (age 23)</td>
<td></td>
<td></td>
<td></td>
<td>5</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Sweep 5 (age 33)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>Sweep 6 (age 42)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5</td>
<td>3</td>
<td>5</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>BM sweep (age 44)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>3</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Sweep 7 (age 46)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Sweep 8 (age 50)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>3</strong></td>
<td><strong>6</strong></td>
<td><strong>5</strong></td>
<td><strong>15</strong></td>
<td><strong>20</strong></td>
<td><strong>17</strong></td>
<td><strong>25</strong></td>
<td><strong>24</strong></td>
<td><strong>27</strong></td>
<td><strong>31</strong></td>
</tr>
</tbody>
</table>
Consistent predictors of participation

• Being female (adulthood only).
• Lower childhood social class in childhood; higher childhood social class in adulthood.
• Higher early life cognitive ability; fewer adolescent conduct problems.
• Social participation; voting; union membership.
• Being married; home ownership.
• Participation in previous sweeps.
Simple test analyses

We test the performance of the missing data strategy using two approaches:

1. “Travelling back in time” to see whether distributions of variables from earlier sweeps can be replicated using only data from respondents at a later sweep.

2. Comparison to external population benchmarks.
Social class of mother’s husband at birth

Percentage in professional social class

4.5

Complete at birth (N = 16,458)

Exclude dead/emigrants (N = 13,880)
Social class of mother’s husband at birth

Percentage in professional social class

- Complete at birth (N = 16,458)
- Exclude dead/emigrants (N = 13,880)
- Respondents age 55 (N = 8284)
Social class of mother’s husband at birth

- Complete at birth (N = 16,458)
- Exclude dead/emigrants (N = 13,880)
- Respondents age 55 (N = 8284)
Social class of mother’s husband at birth

Percentage in professional social class

Complete at birth (N = 16,458)
Exclude dead/emigrants (N = 13,880)
Respondents age 55 (N = 8284)
MI age 55 (N = 13,880)
Multiple imputation

• Complete case analysis only (generally) valid under MCAR.
• Many simple imputation approaches are problematic:
  • Mean imputation.
  • Last observation carried forward.
  • Single conditional imputation.
• In MI, plausible values are used in place of the missing values in a way that allows:
  1. Parameter estimates to be unbiased.
  2. Uncertainty to be estimated in an appropriate way.
• MI valid under MAR.
Multiple imputation

• Specify an appropriate imputation model.
• Create a series of imputed datasets.
• Each imputed dataset analysed using the substantive model.
• Results combined using standard rules.
• Can be undertaken using standard statistical software.
• Widely adopted as practical for applied researchers in a wide range of settings.
Which variables should be included in the imputation model?

Definitely:
• All variables in the substantive model, including any interactions.

Optional “auxiliary variables”:
• Variables associated with the underlying values of the variable(s) subject to missingness.
• Particularly those also associated with the probability of missingness.
Social class of mother’s husband at birth

- Chained equations.
- 50 imputations.
- Auxiliary variables: All predictors of non-response at sweep 9 (age 55) from sweeps 0-8.

1. Complete at birth (N = 16,458)
2. Exclude dead/emigrants (N = 13,880)
3. Respondents age 55 (N = 8,284)
4. MI age 55 (N = 13,880)
Social class of mother’s husband at birth

- Complete at birth (N = 16,458)
- Excluded dead/emigrants (N = 13,880)
- Respondents age 55 (N = 8,284)
- MI age 55 (N = 13,880)
Cognitive ability at age 7

Complete at age 7
(N = 14,407)

Exclude dead/emigrants
(N = 12,938)
Cognitive ability at age 7

- Complete at age 7 (N = 14,407)
- Exclude dead/emigrants (N = 12,938)
- Respondents age 55 (N = 7839)
Cognitive ability at age 7

Mean cognitive ability at age 7

- Complete at age 7 (N = 14,407)
- Exclude dead/emigrants (N = 12,938)
- Respondents age 55 (N = 7839)
Cognitive ability at age 7

- Complete at age 7 (N = 14,407)
- Exclude dead/emigrants (N = 12,938)
- Respondents age 55 (N = 7839)
- MI age 55 (N = 12,938)
Cognitive ability at age 7

- Chained equations.
- 50 imputations.
- Auxiliary variables: All predictors of non-response at sweep 9 (age 55) from sweeps 0-8.
Cognitive ability at age 7

- Complete at age 7 (N = 14,407)
- Exclude dead/emigrants (N = 12,938)
- Respondents age 55 (N = 7839)
- MI age 55 (N = 12,938)
No educational qualifications at age 50

Percentage without educational qualifications

APS GB (N = 673,884)  APS All (N = 796,544)
No educational qualifications at age 50

APS = Annual Population Survey
GB = Born in Great Britain in 1958
All = Born in Great Britain or elsewhere in 1958

APS GB (N = 673,884) APS All (N = 796,544)
No educational qualifications at age 50

Percentage without educational qualifications

- APS GB (N = 673,884)
- APS All (N = 796,544)
- NCDS50 (N = 9783)
No educational qualifications at age 50
No educational qualifications at age 50
No educational qualifications at age 50

- Chained equations.
- 50 imputations.
- Auxiliary variables:
  - MI1: Predictors of educational attainment at age 50.
  - MI2: All predictors of non-response at sweep 8 (age 50) from sweeps 0-7.
  - MI3: All of the above.
No educational qualifications at age 50
No educational qualifications at age 50

Percentage without educational qualifications

- APS GB (N = 673,884)
- APS All (N = 796,544)
- NCDS50 (N = 9783)
- NCDS50 MI1 (N = 15,806)
- NCDS50 MI2 (N = 15,806)
- NCDS50 MI3 (N = 15,806)
No educational qualifications at age 50
More realistic analyses

- Such simple analyses are useful for testing/illustrating the basic idea.
- More realistic analyses likely to be more complicated:
  - More variables in substantive model – exposure(s), outcome(s), control variables,…
  - Inclusion of auxiliary variable(s) predictive of non-response at further sweeps.
  - Inclusion of auxiliary variable(s) predictive of the underlying values of the variable(s) subject to missingness.
  - Inclusion of auxiliary variable(s) to deal with item non-response.
  - Different types of substantive model.
More realistic analyses

• Basic idea remains the same.
• Main concern likely to be (in the MI setting) instability of the imputation model caused by number/type of variables.
• Illustrative realistic example used throughout the NCDS Missing Data User Guide (next session).
Summary

- We have identified variables which predict non-response at each sweep of NCDS.
- 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.
- Will be updated when new sweeps of data become available.
- Also work using linked data (see later).
Handling missing data in the National Child Development Study

User guide (Version 2)

July 2021
1970 British Cohort Study (BCS70)
Identifying predictors of non-response in BCS

• Very similar approach used in BCS 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)
• We also highlight cases in which we can explore MNAR
Identifying predictors of non-response in NCDS

• ~20000 variables in BCS sweeps 0-8.
• 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)
Predictors of non-response

<table>
<thead>
<tr>
<th></th>
<th>NR sweep 1 (age 7)</th>
<th>NR sweep 2 (age 11)</th>
<th>NR sweep 3 (age 16)</th>
<th>NR sweep 4 (age 23)</th>
<th>NR sweep 5 (age 33)</th>
<th>NR sweep 6 (age 42)</th>
<th>NR sweep 7 (age 44)</th>
<th>NR sweep 8 (age 46)</th>
<th>NR sweep 9 (age 50)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sweep 0 (birth)</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>5</td>
<td>1</td>
<td>5</td>
<td>4</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>Sweep 1 (age 7)</td>
<td>4</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Sweep 2 (age 11)</td>
<td></td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Sweep 3 (age 16)</td>
<td></td>
<td></td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Sweep 4 (age 23)</td>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Sweep 5 (age 33)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Sweep 6 (age 42)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sweep 7 (age 46)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sweep 8 (age 46)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>4</td>
<td>7</td>
<td>3</td>
<td>11</td>
<td>7</td>
<td>13</td>
<td>12</td>
<td>8</td>
<td>13</td>
</tr>
</tbody>
</table>
Consistent predictors of participation

- Being female (adulthood only).
- Few household moves
- Paternal social class (early sweeps)
- Higher early life cognitive ability;
- Social participation - voting;
- Home ownership.
- Participation in previous sweeps.
Internal validation: Paternal Social class - at birth

- Chained equations.
- 50 imputations.
- Auxiliary variables: All predictors of non-response at sweep 9 (age 46) from sweeps 0-8.
Internal validation: Cognitive ability - at 7yo

- Complete at 7yo (N=9529)
- Respondents age 46 (N=5097)
Internal validation: Cognitive ability - at 7yo

- Chained equations.
- 50 imputations.
- Auxiliary variables: All predictors of non-response at sweep 9 (age 46) from sweeps 0-8.
Mean BMI levels – Age 34 (MEN)

- Chained equations.
- 50 imputations.
- Auxiliary variables: All predictors of non-response at sweep 6 (age 34) from sweeps 0-5.
SENSITIVITY ANALYSIS – External validation
BMI levels – Age 34 (MEN)

We added a delta value (~0.4kg/m²) in the imputed data in the obs with missing data, after MI, so that MI using delta adj results match the results from HSE.
We followed the same procedure for

- Women at age 34
- Men at age 42
- Women at age 42
Summary

• We have identified variables which predict non-response at each sweep of BCS.
• These can be used as auxiliary variables in subsequent analyses to increase the plausibility of the MAR assumption.
• This approach can be extended for MNAR mechanisms in some cases, with appropriate external benchmark
• Simple test analyses have shown this approach to perform well.
• A straightforward approach, easily implemented in standard software.
Other CLS studies
Next Steps

A data driven approach to understanding and handling non-response in the Next Steps cohort

CLS working paper number 2020/5

By Richard J. Silverwood, Lisa Calderwood, Joseph W Sakshaug, George B. Ploubidis
Millennium Cohort Study (MCS)

- Ongoing work.
COVID-19 Surveys
Non-response weights

- To correct for non-response in the COVID-19 surveys, non-response weights are provided, so that IPW analyses can be undertaken.
- Non-response weights capitalise on the rich data cohort members have provided over many years.
Target population: individuals born in the specified birth period who are alive and still residing in the UK.

Overall response rate within issued sample (39.1%) comparable to similar COVID-19 web surveys.
Derivation of non-response weights

1. Within sample corresponding to target population, model COVID-19 survey response conditional on a common set of covariates using logistic regression.
2. For COVID-19 survey respondents, predict probability of response from model.
3. Calculate non-response weight as inverse of probability of response.
4. Examine distribution of weights across cohorts to decide whether truncation may be desirable; apply truncation if so.
5. Calibrate weights so they sum to number of respondents in each cohort.
Derivation of non-response weights

Response model

- Selection of covariates in response model informed by literature and results of the CLS Missing Data Strategy, plus assumed associations with the probability of response and/or with key COVID-19 survey variables.
- Aimed to use broadly same set of variables in each cohort to ensure consistency.
- Not possible to include identical sets of variables due to data being collected at different ages and using different questions.
### Derivation of non-response weights
#### Response model

<table>
<thead>
<tr>
<th>Sex</th>
<th>Internet access prior to web survey</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ethnicity</td>
<td>Consent for biomarkers</td>
</tr>
<tr>
<td>Parental social class</td>
<td>Consent for linkages</td>
</tr>
<tr>
<td>Number of rooms at home/persons per room</td>
<td>Educational qualifications</td>
</tr>
<tr>
<td>Cognitive ability</td>
<td>Economic activity</td>
</tr>
<tr>
<td>Early life mental health</td>
<td>Partnership status</td>
</tr>
<tr>
<td>Voting</td>
<td>Psychological distress</td>
</tr>
<tr>
<td>Membership in organisations</td>
<td>BMI</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Self-rated health</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smoking status</td>
</tr>
<tr>
<td>Maternal mental health</td>
</tr>
<tr>
<td>Social capital/social support</td>
</tr>
<tr>
<td>Income</td>
</tr>
<tr>
<td>Number of non-responses across all previous sweeps</td>
</tr>
<tr>
<td>Response at COVID-19 Wave 1 survey*</td>
</tr>
</tbody>
</table>
Effectiveness of non-response weights

Grey: all cohort members; red: COVID-19 Wave 2 survey respondents only; blue: COVID-19 Wave 2 survey respondents after application of non-response weights.
Comparison with MI (work in progress)

Grey: all cohort members; red: COVID-19 Wave 2 survey respondents only; blue: after application of non-response weights; green: after application of non-response weights.
Leveraging data linkages
Extending the 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.
- Focusing on linked NCDS and hospital episode statistics (HES) data here. Many other linkages withCLS cohort data available.
Hospital Episode Statistics (HES)

- A collection of databases containing details of interactions with NHS hospitals in England.
- Linkage between NCDS and HES datasets undertaken on the basis of consent at sweep 8 (age 50).
- Matching conducted using deterministic linkage based on combinations of the participant’s name, sex, date of birth and postcode.
- Linked data available via secure access through the UK Data Service.
HES predictors of NCDS non-response

- A total of 58 variables derived from HES data relating to:
  - Numbers of admissions and appointments
  - Missed appointments
  - Investigations undertaken
  - Diagnoses
  - Treatments received
- Employed a similar approach to identify most important predictors of NCDS non-response at wave 9 (age 55).
- 10 variables identified.
Restoring NCDS sample representativeness

- Undertook similar test analysis to see if including the identified HES variables helped restore sample representative.
- Concluded that it did – a bit – but essentially no additional gain relative to using only previously identified survey predictors of non-response.
Using linked Hospital Episode Statistics data to aid the handling of non-response and restore sample representativeness in the 1958 National Child Development Study

GLS working paper number 2022/1

Nasir Rajah1, Lisa Calderwood1, Bianca L De Stavola2, Katie Harron2, George B Ploubidis1 and Richard J Silverwood1

1. Centre for Longitudinal Studies, UCL Social Research Institute, 20 Bedford Way, London WC1H 9AL
2. Population, Policy & Practice Research and Teaching Department, UCL Great Ormond Street Institute of Child Health, 30 Guilford Street, London WC1N 1EH
Resources
Resources

• Handling missing data webpage: https://cls.ucl.ac.uk/data-access-training/handling-missing-data/
Resources


- 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.

Resources


Thank you.