Getting started with the data
This section

- Available resources
- Accessing the data
- Key ID’s and other data protocols
- Merging data within and across sweeps
- Study design and sample weights
- Non-response and attrition
- Where to go for more information
Available resources  https://cls.ucl.ac.uk/

- User guides
  - Overview of measures
  - Response and weights

- Questionnaires
  - Exact question wording
  - Questionnaire routing
  - Variable names

- Data documentation
  - Data notes
  - Coding frames
  - Variables lists, including derived variables

- Technical reports
  - Sample and questionnaire design, development
  - Fieldwork, response, ethics
  - Coding, editing

- Data dictionaries

- Cohort profiles e.g.

- Previous journal publications
Resources available: UK Data Service
Data freely available to researchers, government analysts and third sector
Access to different types of data

Access to data held by the UK Data Service varies depending on how the data is classified:

Tier 1: End User Licence (EUL) for access to data with a low level of sensitivity and disclosivity.
• Most of our data are available under this licence.
• Your application is authorised directly by the UK Data Service, and you can download the data directly from there.

Tier 2a: Special Licence (SL) for access to moderately sensitive or disclosive data. Access through the UK Data Service and application approved by CLS before you can download the data.

Tier 2b: Secure Access Licence (SA) for access to the most sensitive and/or potentially disclosive data. Access through the UK Data Service and attend a specialised training course.

https://cls.ucl.ac.uk/data-access-training/data-access/
1. Click 'Login/Register'
2. On the Login page begin typing your organisation name. Select the organisation required
3. Click 'Continue'
4. Your own organisation login page will then be displayed. Login with your usual username and password
5. Complete the registration form with your details, selecting other options as required
6. Agree to the End User Licence (EUL), which outlines the terms and conditions of use of the Service
7. Click 'Register'.

Once registration is complete you will be able to download/order or request access to data
Accessing CLS cohort data

<table>
<thead>
<tr>
<th>Cohort</th>
<th>Link</th>
</tr>
</thead>
<tbody>
<tr>
<td>NCDS</td>
<td><a href="http://discover.ukdataservice.ac.uk/series/?sn=2000032">http://discover.ukdataservice.ac.uk/series/?sn=2000032</a></td>
</tr>
<tr>
<td>BCS70</td>
<td><a href="http://discover.ukdataservice.ac.uk/series/?sn=200001">http://discover.ukdataservice.ac.uk/series/?sn=200001</a></td>
</tr>
<tr>
<td>Next Steps</td>
<td><a href="http://discover.ukdataservice.ac.uk/series/?sn=2000030">http://discover.ukdataservice.ac.uk/series/?sn=2000030</a></td>
</tr>
<tr>
<td>MCS</td>
<td><a href="http://discover.ukdataservice.ac.uk/series/?sn=2000031">http://discover.ukdataservice.ac.uk/series/?sn=2000031</a></td>
</tr>
</tbody>
</table>

**Millennium Cohort Study**

Abstract

The Millennium Cohort Study (MCS), which began in 2000, is conducted by the Centre for Longitudinal Studies (CLS). It aims to track the conditions of social, economic and health advantage and disadvantage facing children born at the start of the 21st century. The study has been following the 'Millennium children' through their early childhood years and plans to follow them into adulthood. It aims to provide a basis for comparing patterns of development with the preceding cohort studies (the National Child Development Study (NCDS) and the 1970 Birth Cohort Study (BCS70)).

Access data

[Access Millennium Cohort Study - Survey and Biomessures Data]

[Access Millennium Cohort Study - Linked Administrative Data]
Accessing CLS cohort data II

Before downloading the data:
- Click on Request Access
- Click on Complete actions
- Agree to standard ‘End User Licence’
- Read and agree extra conditions

Choose data format and download zip file
- SPSS
- STATA
- TAB (tab-delimited)
<table>
<thead>
<tr>
<th>Name</th>
<th>Contents</th>
<th>Structure</th>
<th>Identifier</th>
</tr>
</thead>
<tbody>
<tr>
<td>NS8_2015_Main_Interview</td>
<td>Modules 1 to 7</td>
<td>Flat</td>
<td>NSID</td>
</tr>
<tr>
<td>NS8_2015_Self_Completion</td>
<td>Module 8</td>
<td>Flat</td>
<td>NSID</td>
</tr>
<tr>
<td>NS8_2015_Partnerships</td>
<td>Relationship histories</td>
<td>Hierarchical</td>
<td>NSID, W8RELID</td>
</tr>
<tr>
<td>NS8_2015_Children</td>
<td>Details of children of CM</td>
<td>Hierarchical</td>
<td>NSID, W8CHID</td>
</tr>
<tr>
<td>NS8_2015_Household_Members</td>
<td>Details of members living in same household as CM</td>
<td>Hierarchical</td>
<td>NSID, W8HHMID</td>
</tr>
<tr>
<td>NS8_2015_Activity_History</td>
<td>Activities and Employment histories</td>
<td>Hierarchical</td>
<td>NSID, W8HISTID</td>
</tr>
<tr>
<td>NS8_2015_Benefits</td>
<td>Details of individual benefits received</td>
<td>Hierarchical</td>
<td>NSID, W8BENID</td>
</tr>
<tr>
<td>NS8_2015_Income_Unfolding_brackets</td>
<td>Unfolding brackets questions for payments and income</td>
<td>Flat</td>
<td>NSID</td>
</tr>
<tr>
<td>NS8_2015_Benefits_Unfolding_brackets</td>
<td>Unfolding brackets questions for benefits</td>
<td>Hierarchical</td>
<td>NSID, W8BENID</td>
</tr>
<tr>
<td>NS8_2015_Derived_variables</td>
<td>Derived variables</td>
<td>Flat</td>
<td>NSID</td>
</tr>
</tbody>
</table>
## Key identifiers (ID’s)

<table>
<thead>
<tr>
<th>Cohort</th>
<th>Key cohort identifier</th>
<th>Key cohort member/family identifier format</th>
</tr>
</thead>
<tbody>
<tr>
<td>NCDS 1958</td>
<td>NCDSID</td>
<td>7 characters:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>N followed by 5 digits, and a single character</td>
</tr>
<tr>
<td></td>
<td></td>
<td>e.g. N10016V</td>
</tr>
<tr>
<td>BCS70 1970</td>
<td>BCSID</td>
<td>7 characters:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B followed by 5 digits, and a single character</td>
</tr>
<tr>
<td></td>
<td></td>
<td>e.g. B25819Z</td>
</tr>
<tr>
<td>Next Steps 1989/90</td>
<td>NSID</td>
<td>8 characters:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NS followed by 5 digits and a single character</td>
</tr>
<tr>
<td></td>
<td></td>
<td>e.g. NS21140C</td>
</tr>
<tr>
<td>MCS 2000/02</td>
<td>MCSID</td>
<td>7 characters:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>M followed by 5 digits, and a single character</td>
</tr>
<tr>
<td></td>
<td></td>
<td>e.g. M10029A</td>
</tr>
</tbody>
</table>

### Cohort member/family identifier:

Every cohort member (or for the MCS family) has the same ID across sweeps.

Use these ID’s to link datasets.

---

**W8xxID** used in particular files to denote relationship, child, HH member etc.

**CNUM** Cohort members, 1, 2 (twins) or 3 (triplets)

**PNUM** Person number, for everyone else in the family apart from cohort members: parents, siblings, grandparents, etc.
# File structures: Flat v hierarchical

<table>
<thead>
<tr>
<th>Type of file structure</th>
<th>Format</th>
<th>Examples in the cohorts</th>
</tr>
</thead>
</table>
| Flat                   | 1 record per case | NCDS, BCS70, Next Steps – main data files  
MCS – family files |
| Hierarchical           | 1 or more records per case | Household files  
Activities e.g. employment histories  
Relationship histories  
Time use diaries  
MCS - Person within family |

<table>
<thead>
<tr>
<th>BCSID</th>
<th>Sex</th>
<th>Country</th>
<th>Emp1</th>
<th>Emp2</th>
<th>Emp3</th>
</tr>
</thead>
<tbody>
<tr>
<td>B567689A</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>B567689A</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>B567689A</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>B467921B</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B879255C</td>
<td>2</td>
<td>2</td>
<td>5</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>B879255C</td>
<td>2</td>
<td>2</td>
<td>5</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>B297614D</td>
<td>1</td>
<td>1</td>
<td>6</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>B297614D</td>
<td>1</td>
<td>1</td>
<td>6</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>B349725E</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Merging data within and across sweeps

1. Identify appropriate files
   - Establish number of cases in target population

2. Check file structure: flat v hierarchical
   - Transform if necessary

3. Identify merging variables:
   1. Unique ‘key’ cohort ID (member or family)
   2. Other ID’s depending on merge
      • Check the same variable name (case sensitive, changed across sweep etc)
      • Create identical variable name if necessary

4. Check merged correctly
## Study design (sampling and sample weights)

Studies are representative of...

<table>
<thead>
<tr>
<th>Cohort</th>
<th>Population</th>
<th>Sample and study design</th>
<th>Design weight(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NCDS 1958</td>
<td>All born in GB in one week</td>
<td><strong>Total sample:</strong> Captured 98% of the total births in GB in the target week</td>
<td>None</td>
</tr>
<tr>
<td>BCS70 1970</td>
<td>All born in GB in one week</td>
<td><strong>Total sample:</strong> Captured 95-98% of the total births in GB in the target week</td>
<td>None</td>
</tr>
<tr>
<td>Next Steps 1989/90</td>
<td>Young people in England in 2004 born between 01/09/89 and 31/8/90</td>
<td><strong>Complex sample design:</strong> Maintained schools* 2-stage sampling procedure</td>
<td>psu (SampPSU) strata (SampStratum)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Stage 1: Schools</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Stage 2: Pupils within schools.</td>
<td></td>
</tr>
<tr>
<td>MCS 2000/02</td>
<td>Children born in 2000 /2002 and living in the UK at age nine months,</td>
<td><strong>Complex sample design:</strong> Clusters: Areas of residence (electoral wards)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Disproportionately stratified by area disadvantage, and UK country, and ethnicity in England</td>
<td></td>
</tr>
</tbody>
</table>

*Independent and referral units sampled differently*
Non-response and attrition

- Distinction between unit (respondents’) non-response and item non-response
  - Unit non-response (not responding to a particular sweep)
    - Non-response is common in longitudinal surveys
  - Item non-response i.e. not answering some questions
    - Tends to be less of an issue in the cohorts

- Missing data may be a risk to representativeness
  - Potential for bias since respondents are often systematically different from nonrespondents
Dealing with unit non-response

- Case-wise deletion i.e. ignoring non-response (unless missing completely at random)
  - Any individual in a data set is deleted from an analysis if they’re missing data on any variable in the analysis
  - Straightforward, but doesn’t deal with any non-response bias

- Non-response weights
  - Adjust the sample composition to take account of the loss of particular type of respondents.
  - Provided in MCS (govwt2 = overall in MCS7 for whole of UK analysis)
  - Provided in Next Steps (W8FINWT = final weight for age 25 survey)
  - Provided in User guide (Version 2)

- Other more advanced methods e.g. multiple imputation (MI)
  - MI involves the generation of multiple copies of the dataset in each of which missing values are replaced by imputed values sampled from their posterior predictive distribution given the observed data.
Where to go for more information
CLOSER Learning Hub: [https://learning.closer.ac.uk/](https://learning.closer.ac.uk/)

CLOSER provides training and resources for students and early-career researchers to “maximise the use, value and impact of longitudinal research”

Learning Hub – demonstration video: [https://youtu.be/Z_bFCClq2Dc](https://youtu.be/Z_bFCClq2Dc)
Resources

NCRR has many online resources that are intended to help people interested in social science research methods. Here is a selection of the current NCRR resources.

Courses and events
Publications database - NCRR EPrints
Audio podcasts
Videos
Online learning resources
Resources for trainers
COVID-19 resource collection

https://www.ncrm.ac.uk/resources/
Resources available: UK Data Service learning hub

https://ukdataservice.ac.uk/learning-hub/
Resources available: CLS website

https://cls.ucl.ac.uk/

COVID-19 survey

During the coronavirus pandemic, we have been running a series of surveys to find out about the experiences of the participants in five national longitudinal cohort studies. The aim is to understand the economic, social and health impacts of the COVID-19 crisis, the extent to which the pandemic is widening or narrowing inequalities, and the lifelong factors which shape vulnerability and resilience to its effects.

We have now completed three waves of the survey. Participants in all four of the national longitudinal cohort studies that we manage at CLS, as well as participants in the MRC National Survey of Health and Development, have taken part. These studies have been following large nationally representative groups of people since birth, with ages currently ranging from 13 through to 74:

- Millennium Cohort Study (born 2000–02),
- Next Steps (born 1989–90),
- BCS70 (born 1970),
- 1946 Birth Cohort (born 1946),
- 1958 Birth Cohort (born 1958),

BCS70 sweeps

Since the birth survey in 1970 there have been five ‘sweeps’ of all cohort members. Click on a sweep below to learn more about the information collected. The latest sweep, at age 51, is now underway.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>Birth</td>
<td>5</td>
<td>10</td>
<td>15</td>
<td>20</td>
<td>25</td>
<td>30</td>
<td>35</td>
<td>40</td>
<td>45</td>
<td>50</td>
<td>51</td>
</tr>
</tbody>
</table>

On this page: Introduction  Sweeps  COVID-19 survey and data  Sub studies  50 years in 50 weeks  Latest from BCS70  Report publications  Study features  Popular documentation  Data access  Principal Investigator  More related content
Resources available on each cohort and sweeps:
User Guide, technical resources and questionnaires
CLS training and support

Upcoming training events

<table>
<thead>
<tr>
<th>Event</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Using health administrative data: Health Episode Statistics</td>
<td>6 December 2022</td>
</tr>
<tr>
<td>Introduction to the National Child Development Study (NCDS)</td>
<td>25 January 2023</td>
</tr>
<tr>
<td>Handling Missing Data in the cohorts (NCDS)</td>
<td>28 February 2023</td>
</tr>
</tbody>
</table>

https://cls.ucl.ac.uk/data-access-training/training-and-support-2/
So we’ve covered

- An introduction to birth cohorts
- Some of the content in the CLS cohorts by subject areas
- Examples of the types of analysis
- Getting started with the data
- Where to go for more information
Thank you
Any questions?