Institute of Education

UCL

Millennium Cohort Study

Linked education administrative datasets KS1-KS4 – Wales

User Guide (Version 2)

October 2023

CENTRE FOR LONGITUDINAL STUDIES

UKRI Economic and Social Research Council

Contact

Data queries: help@ukdataservice.ac.uk

Questions and feedback about this user guide: clsfeedback@ucl.ac.uk.

Authors

Karen Dennison, Andrew Peters, Emla Fitzsimons

How to cite this guide

Dennison, K., Fitzsimons, E., Peters, A. (2022) *Millennium Cohort Study: A guide to the linked education administrative datasets – Wales. User Guide (Version 1).* London: UCL Centre for Longitudinal Studies.

You should also acknowledge CLS following the guidance from https://cls.ucl.ac.uk/data-access-training/citing-our-data/

This guide was published in March 2023 by the UCL Centre for Longitudinal Studies.

Centre for Longitudinal Studies (CLS)
UCL Social Research Institute
University College London
20 Bedford Way, London WC1H 0AL

Centre for Longitudinal Studies

The UCL Centre for Longitudinal Studies (CLS) is an Economic and Social Research Council (ESRC) Resource Centre based at the UCL Social Research Institute, University College London. It manages four internationally-renowned cohort studies: the 1958 National Child Development Study, the 1970 British Cohort Study, Next Steps, and the Millennium Cohort Study. For more information, visit www.cls.ucl.ac.uk.

This document is available in alternative formats. Please contact the Centre for Longitudinal Studies:

tel: +44 (0)20 7612 6875
email: clsfeedback@ucl.ac.uk

Contents

About the Millennium Cohort Study

The Millennium Cohort Study (MCS) is a longitudinal birth cohort study, following a nationally representative sample of approximately 19,000 people born in the UK at the turn of the century.

The study has captured rich information about the different aspects of cohort members' lives, from birth to childhood and adolescence, and is continuing to keep up with them now they are adults.

As a multidisciplinary study, MCS is used by researchers working in a wide range of fields. Findings from MCS have influenced policy at the highest level, and today the study remains a vital source of evidence on the major issues affecting young people's lives.

Further details of the data available from the main surveys can be found on the CLS website www.cls.ucl.ac.uk/cls-studies/millennium-cohort-study/ and, in particular, the MCS Guide to the Datasets https://cls.ucl.ac.uk/data-access-training/exploring-our-data/

1. Introduction

This guide describes the data linkage of education administrative records for Wales up to age 16 to survey data for cohort members in the Millennium Cohort Study (MCS). The main aim of this data linkage exercise is to enhance the research potential of the study, by combining administrative education records with the rich information collected in the surveys.

2. Consent to education data linkage

During the fourth (age 7) sweep, the child's parent / carer was asked for consent to link the child's education records up to age 16 to their MCS survey data. During the seventh sweep (age 17) the cohort members, then able to consent in their own right, were asked for consent to link to post-16 education records.

Cohort members can withdraw their or their parents' consent for further linkages at any time. Detailed information on the fieldwork and consent collection can be found in the MCS Age 6 and MCS Age 17 Technical reports and their appendices. All documents can be found under 'documentation' at:

- cls.ucl.ac.uk/cls-studies/millennium-cohort-study/mcs-age-7-sweep/
- cls.ucl.ac.uk/cls-studies/millennium-cohort-study/mcs-age-17-sweep/

3. Education data linkage

3.1 Matching strategy

Welsh Government schools and pupil data are held by the SAIL Databank. In 2019 CLS provided NHS Wales Informatics Service (NWIS - now called Digital Health Care Wales) with the direct identifiers of those MCS cohort members whose parents had consented to education linkages up to age 16 and for whom consent had not been subsequently withdrawn.

NWIS matched those records to those held in the Welsh Demographic Service, anonymised them and assigned each one a unique, non-identifiable code. NWIS then sent this code, and minimal information on gender, area of residence and week of birth to SAIL Databank so that the data could be matched to the education records held in the SAIL Databank.

Researchers can apply to access the education data linked to MCS survey responses through the SAIL Databank.

In 2020 CLS also received approval from the SAIL Databank to extract the linked education data for data sharing via the UK Data Service. CLS received the extracted files in 2021.

3.2 Matching rates

During the fourth (age 7, 2008) sweep the parents of 13,170 out of 14,043 cohort members consented to education linkages, a 94% consent rate. Out of these, 1,978 cohort members in Wales participated in the study. Of the 13,165 records,1,922 had an interview address in Wales in MCS4, MCS5, MCS6 or for current/last known address (as in 2019).

In 2019, CLS sent NWIS a matching file containing the information from the full consenting cohort regardless of whether their MCS interview address had ever been in Wales to match as many records as possible. 13,165 cohort members who still had valid consents. The file included surname, forenames, addresses/postcodes (for MCS4, MCS5, MCS6 and current / last known address in 2019), date of birth and gender.

A total of 1,940 MCS cohort members were successfully matched to their Welsh education records. This figure is higher than the number of cohort members with interview addresses in Wales – these additional cases are likely to be people who weren’t interviewed in Wales but who moved to/from Wales between interviews.

Table 1 below shows the number of successful matches to education records following data linkage.

Table 1. Consent and overall linkage

Number in MCS Age 7 Survey	14,043
Number with valid consent	13,170
Consent rate	94%
Total number sent for matching	13,165
Number of participants in Wales in 2008	1,978
Total number of consenting participants who had interview address in Wales in MCS4, MCS5, MCS6 or 2019	1,922
Consent rate for participants with interview address in Wales	96%
Total number with matched education data (prior to withdrawals)	1,940
Linkage rate	100%

Data is available and matched for a total 1,922 cohort members across a number of different tables.

4. Linked Research Data

4.1 Licencing and data access

The linked education data are available as controlled data from the UK Data Service (UKDS) SecureLab.

All users of the data need to be registered with the UKDS. Details of how to do this are available at https://www.ukdataservice.ac.uk/get-data/how-to-access/registration.

Access to the UKDS Secure Lab can take place via the researcher's own institutional desktop PC, using a Safe Pod or at the Safe Room at the UK Data Archive. Applicants wishing to access this data need to abide by the terms and conditions of the UKDS Secure Access licence. Before gaining access, researchers must make an

application detailing the intended analysis and provide a justification as to why this data is requested. Application guidance can be found at https://ukdataservice.ac.uk/find-data/access-conditions/secure-application-requirements/apply-to-access-non-ons-data/.

4.2 Datasets

The linked education data available at the UKDS Secure Lab included 15 datasets, each covering different areas and periods of education described in Table 2:

Table 2. Description of datasets

Dataset name	Description	Number of cohort members
mcs_eduw_attendance	Possible sessions and number of sessions not attended (authorised/unauthorised absences shown separately)	1,859
mcs_eduw_eotas_provision	Education provision details for Educated Other Than At School (EOTAS) children	24
mcs_eduw_eotas_pupil	Pupil details for children Educated Other Than At School (EOTAS)	53
mcs_eduw_eotas_sen	Details for Educated Other Than At School (EOTAS) children with Special Educational Needs (SEN)	24
mcs_eduw_exclusions	Exclusions of children in state schools	196
mcs_educ_hiru_plasc_pupil	Pupil Level Annual School Census (PLASC) records gathered mainly from the annual school census	1,885
mcs_eduw_census_pupil	Pupil census based on an annual snapshot	1,891
mcs_eduw_plasc_sen	Pupil Level Annual School Census (PLASC) records of children with Special Educational Needs (SEN)	1,866
mcs_educ_hiru_ks123	Teaching assessment for Key stages 1,2 and 3	1,851
mcs_educ_pre16_ks1	Pre-16 results Foundation phase	1,729
mcs_eduw_ks1	KeyStage 1 results Foundation phase	1,735

mcs_eduw_ks2	KeyStage 2 results Year 6	1,793
mcs_eduw_ks3	KeyStage 3 results Year 9	1,765
mcs_eduw_ndc_pupil	Results of the end of Foundation Phase and Key Stage 2 and 3	1,813
mcs_eduw_wed_l1	KeyStage 4 results (Welsh examinations database)	1,762

Additional details on how the data are structured can be found at Education Wales (EDUW) (healthdatagateway.org).

4.3 Identifiers

MCS data are identified with the same research IDs used for the rest of cohort data available at the UKDS. This enables the data to be easily merged with one another.

For MCS, researchers need to use both the MCS family identifier (MCSID) and the two individual person identifiers (CNUM00/PNUM00) to merge on with other cohort data. As CNUM00 and PNUM00 include the wave number they may need consistent naming across datasets beforehand depending on the method of merging used.

The majority of datasets are hierarchical (long format). Secondary identifiers are ordered by calendar year, though the order within each year is not chronological as more precise dates are not available. For example, where a pupil had multiple school absences in the same calendar year the first ordered absence may not be the earliest to occur.

There are different ways the data of MCS can be merged depending on the focus of the research project (Parent/Carers, Cohort Members or family). Details, syntax and examples on merging is provided by the MCS Data Handling Guide.

4.4 Data processing

Variable names

With the exception of individual identifiers, the variable names are those provided by the SAIL Databank. and typically give some indication of the variable contents.

Variable labels and value labels

Variable labels and value labels are all provided by the SAIL Databank. The majority of the data was initially text-based, with values represented by single characters or abbreviations. For the sake of data usability these have been converted to numeric values and the appropriate value labels attached. In some circumstances value labels did not exist for the original text-based value, when this occurred it was these abbreviations that were given numeric values and set as the value label.

As such it may be apparent in the data that some labels are unclear, or lacking quality compared to others. However, it was decided these are better to be included as they are rather than to set the values to missing should users want to infer their meaning.

Missing data

For all empty fields a value of -8 (no information) was assigned, indicating no data was (or could be) collected. There are no other missing values.

4.5 Data de-identification

CLS is committed to protect research participants' rights and avoid data disclosure and re-identification of individuals. As such, all instances of Local Education Authority (LEA) codes across the data have been replaced by anonymised 4-digit 'anonlea' codes. This allows data users to identify which pupils are in the same LEA, or when and if they changed LEA, without their geographical location being disclosed.

5. Disclosure control: UK Data Service (UKDS) requirements for data users

As the education data linked to the longitudinal MCS data are only available via the UKDS Secure Lab, the UKDS will always perform a certain level of disclosure control on the outputs generated by researchers, as outlined in their SDC Handbook, which can be downloaded from https://securedatagroup.org/sdc-handbook/.

The two UKDS Secure Lab rules of thumb that will be applied to all outputs are:

- Threshold rule: No cells should contain less than 10 observations;

- Dominance rule: No observation should dominate the data to a huge extent.