IOE, Faculty of Education and Society

UCL

Fertility histories in four British cohort studies

User Guide (Version 1)

June 2025

CENTRE FOR LONGITUDINAL STUDIES

UKRI Economic and Social Research Council

Contact

Data queries: help@ukdataservice.ac.uk

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

Authors

Aase Villadsen, Samantha Parsons, Alice Goisis

How to cite this guide

Villadsen A., Parsons S., Goisis A. (2025) *Fertility histories in four UK cohort studies: User Guide (Version 1)*. UCL Centre for Longitudinal Studies.

Data citation and acknowledgement

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

Centre for Longitudinal Studies

Centre for Longitudinal Studies (CLS)

UCL Social Research Institute

University College London

20 Bedford Way, London WC1H 0AL

www.cls.ucl.ac.uk

The UCL Centre for Longitudinal Studies (CLS) is an Economic and Social Research Council (ESRC) Resource Centre. It is home to a unique series of UK national cohort studies. It is part of the UCL Social Research Institute, based at the IOE, UCL's Faculty of Education and Society.

This document is available in alternative formats. Please email the Centre for Longitudinal Studies at clsdata@uck.ac.uk

Contents

1. Introduction

This data guide has been developed to accompany the deposit of datasets on the Cohort Members (CM) fertility histories in four UK cohort studies. The cohorts included are:

- MRC National Survey of Health and Development (born 1946)
- National Child Development Study (born 1958)
- British Cohort Study (born 1970)
- Next Steps (born 1989/90).

The aim is to enhance data from these four British birth cohorts born between 1946 and 1990, through retrospective harmonisation of information on the CM fertility which serves a dual purpose. First, it enables researchers to improve the measurement of fertility (as an outcome, predictor, or control variable) within cohort analyses. Second, it facilitates cross-cohort research on fertility using these rich datasets.

We created a separate longitudinal dataset for each cohort, with data covering survey sweeps from when cohort members were young adults in their early to mid-twenties, until early fifties which for most marks the end of the reproductive window. The derived variables provide a summary of fertility (live births) for each cohort member at each survey sweep, such as whether the cohort member had had any children, number of children, age of the eldest and youngest child, and number of boys and girls. The focus was on live births rather than pregnancies that were terminated, miscarriages, or stillbirths.

Shown in Box 1 is the full list of target variables. Due to lack of information, it was not possible to derive all target variables, at all sweeps, in all cohorts. The NSHD especially was affected by this, whereas a larger number of harmonised variables could be derived in the NCDS, BCS70, and Next Steps, which share many identical survey questions. The sections that follow outline narratively the derivation approach taken in each cohort. In addition, annotated Stata code is available, which allows data users to inspect the derivation process, as well as adapt the code for derivation

of further and alternative fertility variables (links are provided further below for each cohort).

Box 1.1: List of target fertility variables for Cohort Members (CM)

CM Biological children

- Whether has had any bio children
- Number of bio children
- Flag: More biological children reported at previous age than at current age
- Number of bio children in HH
- Flag: More bio children reported in HH grid than in pregnancy data
- Number of bio children not in HH
- Number of bio children had with a previous partner
- Have had any bio children with a previous partner
- Age in years of eldest bio child
- Age in years of youngest bio child
- Age in years of CM at birth of eldest bio child
- Age in years of CM at birth of youngest bio child
- Number of bio children who are boys
- Number of bio children who are girls

CM Non-biological children

- Whether has any non-bio children in HH
- Number of non-bio children in HH
- Number of adopted children in HH
- Number of foster children in HH
- Number of stepchildren in HH
- Age in years of eldest non-bio child
- Age in years of youngest non-bio child
- Number of non-bio children who are boys
- Number of non-bio children who are girls

CM Biological or non-biological children

- Whether has any children (bio or non-bio)
- Number of children (bio or non-bio)
- Age in years of eldest child (bio or non-bio)
- Age in years of youngest child (bio or non-bio)
- Number of children who are boys (bio or non-bio)
- Number of children who are girls (bio or non-bio)

CM Partner and/or children

- Whether has a partner in HH
- Marital status
- Whether has live-in partner/spouse partner and/or any bio children
- Whether has live-in partner/spouse partner and/or any bio or non-bio children

Other variables

- CM Research case identifier
- Sex of CM
- Birth year of CM

- Birth month of CM
- Whether CM took part in survey sweep
- Interview year
- Interview month
- Survey weights (in applicable cohorts)

2. Fertility histories in 1946 MRC National Survey of Health and Development (NSHD)

2.1 About the NSHD

The MRC National Survey of Health and Development (NSHD) study is the oldest and longest running of the British birth cohort studies and is funded by the Medical Research Council (MRC). From an initial maternity survey of 13,687 of all births recorded in England, Scotland and Wales during one week in 1946, a socially stratified sample of 5,362[1] singleton babies born to married parents was selected for follow-up. This sample has been studied 25 times, with research findings having informed UK health care, education and social policy for more than 70 years. During their childhood, the main aim of the NSHD was to investigate how the environment at home and at school affected physical and mental development and educational attainment. During adulthood, the main aim was to investigate how childhood health and development and lifetime social circumstances affected their adult health and function and how these change with age. Today, with study members in their seventies, the NSHD offers a unique opportunity to explore the long-term biological and social processes of ageing and how ageing is affected by factors acting across the whole of life. For more information on the NSHD visit their website.

The NSHD has a data archive with over 35,000 variables collected over the lifetime of the study. The archive preserves data and meta-data (descriptions of the data) in electronic, paper, fiche, and, increasingly, image form, and includes biological samples. The NSHD now has a set of persistent identifiers that allow the tracking of

[1] There are n=5,360 study members in this dataset.

research datasets as well as publications. The DOI for data collected from 1946-2005 is: 10.5522/NSHD/Q101

2.2 Original data that were used from the NSHD

Data on fertility were drawn from sweeps when study members were aged 19, 20, 22, 26, 31, 36, 43 and 53, which were collected between 1965 and 1999. To access the raw MRC NSHD data, a request was made to the NSHD Data Sharing Committee (see section 6.2 below for details). Links to the specific questionnaires that contain the information on fertility and partnerships are included in Table 2.1.

Table 2.1: Overview of original data used for derivation of fertility variables at each sweep in the MRC NSHD

Sweep	Data sections used for derivation	Sample size
Age 19 (Sweep x, 1965)	Self-Completion Questionnaire	3,559
Age 20 (Sweep x, 1966)	Self-Completion Postal Questionnaire	3,897
Age 22 (Sweep x, 1968)	Self-Completion Postal Questionnaire	3,883
Age 26 (Sweep x, 1972)	Main Questionnaire – Section A	3,749
Age 31 (Sweep x, 1977)	Self-Completion Postal Questionnaire	3,339
Age 36 (Sweep x, 1982)	Main Questionnaire – Form A	3,321
Age 43 (Sweep x, 1989)	Main Questionnaire – Form A	3,261
Age 53 (Sweep x, 1999)	Main CAPI Questionnaire – Form A	3,034

2.3 How fertility variables were derived in NSHD

The ways in which a study members fertility was captured varied across the sweeps, depending on the information collected. In the earlier sweeps (age 19, 20 and 22), study members were asked to report if they had a child and/or the number of children they had at each sweep. This varied between one or two questions. At age 26, the age they had each child, and the sex of each child was collected for the first

time. At age 31 and 43, this was repeated for any additional children they had since their last interview, or for study members not participating at an earlier age, details on the age they had each of all their children and the sex of each child was collected. At age 36, information on the age they had each child, and the sex of each child was again collected from all participating study members. Information on step, fostered or adopted children was collected at age 36 and 43. At age 53 any further children a study member had was collected, together with the sex of the child and the year they were born.

Annotated Stata code is available here for further inspection of the variable derivation process in the NSHD.

2.4 Data errors and inconsistencies in NSHD

Given study members were asked questions about their fertility at numerous sweeps and data collection was not computerised until the age 53 sweep in 1999, some data inconsistencies are to be expected. A 'flag' variable has been derived to identify mismatch information over time, such as when a child was mentioned at an earlier sweep but not at a subsequent sweep, or when the total number of children was less than had been mentioned at an earlier sweep.

2.5 Survey and non-response weights in NSHD

When analysing the 1946 data, users need to weight the data to adjust the results for the socially stratified sample that was followed up after the original birth survey.

Example Stata code to apply the weight

```
proportion anybiochildren [pw=inf], over(sex)
```

3. Fertility histories in 1958 National Child Development Study (NCDS)

3.1 About the NCDS

The 1958 National Child Development Study (NCDS) is an ongoing birth cohort study, that has followed the lives of over 17,000 people born in England, Scotland, and Wales in a single week in March of 1958. The first survey was carried out at the birth of the child, with subsequent follow up surveys at age 7, 11, 16, 23, 33, 42, 44, 46, 50, 55, and 62 years. The initial survey was known as the 'Perinatal Mortality Study' with a focus on factors associated with the health of mothers and their newborn babies. Over time the study has become increasingly multidisciplinary with the collection of many different of aspects of participants lives across their life course. Further information about the NCDS can be found on the website for the Centre for Longitudinal Study (CLS).

3.2 Original data that were used from the NCDS

Data on fertility were drawn from age 23, 33, 42, 46, and 50 (sweep 4 to 8), which were collected between 1981 and 2008. These original data are available under End User Licence from the UK Data Service.

Citation for the NCDS datasets is:

University College London, UCL Social Research Institute, Centre for Longitudinal Studies. (2024). National Child Development Study. [data series]. 14th Release. UK Data Service. SN: 2000032, DOI: http://doi.org/10.5255/UKDA-Series-2000032

Table 3.1: Overview of original data used for derivation of fertility variables at each sweep in the NCDS

Sweep	Data sections used for derivation [data file name]	Sample size
Age 23 (Sweep 4, 1981) Survey mode: Face to face	-Main questionnaire *[*ncds4*]:* marital status and cohabitation, interview date. -Child data *[*ncds4*]:* any children and number, whether alive, sex, and month and year of birth -Household grid *[*ncds4*]:* sex, age and relationship to CM of all persons living in the household at sweep.	12,537
Age 33 (Sweep 5, 1991) Survey mode: Face to face with self-completion questionnaire	-Main questionnaire *[ncds5cmi]*: marital status and cohabitation, interview date. -Pregnancy data *[ncds5cmi]:* all pregnancies ever, outcome of pregnancy, sex, month and year of birth, where child is now, other parent of child. -Household grid *[ncds5cmi]:* sex, age and relationship to CM of all persons living in the household at sweep.	11,469
Age 42 (Sweep 6, 2000) Survey mode: Face to face	-Main questionnaire *[ncds6]:* marital status and cohabitation, interview date. -Pregnancy data *[ncds6]:* new pregnancies since last interview, outcome of pregnancy, sex, month and year of birth. -Household grid *[ncds6]:* sex, age and relationship to CM of all persons living in the household at sweep.	11,419
Age 46 (Sweep 7, 2004) Survey mode: Telephone	-Main questionnaire *[ncds7]:* marital status and cohabitation, interview date. -Pregnancy data *[ncds7]:* new pregnancies since last interview, outcome of pregnancy, sex, month and year of birth. -Household grid *[ncds7]:* sex, age, and relationship to CM of all persons living in the household at sweep.	9,534
Age 50 (Sweep 8, 2008)	-Main questionnaire *[ncds_2008_followup]:* marital status and cohabitation, interview date.	9,790

Sweep	Data sections used for derivation [data file name]	Sample size
Survey mode: Face to face plus self-completion questionnaire	-Pregnancy data *[ncds_2008_followup]:* new pregnancies since last interview, outcome of pregnancy, sex, month and year of birth. -Household grid *[ncds_2008_followup]:* sex, age, and relationship to CM of all persons living in the household at sweep.	

Guides to the datasets for each NCDS sweep are available from the CLS website via the following weblinks:

Sweep 4, Age 23, 1981

Sweep 5, Age 33, 1991

Sweep 6, Age 42, 2000

Sweep 7, Age 46, 2004

Sweep 8, Age 50, 2008

3.3 How fertility variables were derived in NCDS

Fertility variables for biological children were mainly derived from pregnancy data where cohort members reported outcomes of conceived or fathered pregnancies. For each live birth, information was collected on the child's sex, month and year of birth, and whether the child lived with the respondent. At age 23 and 33, cohort remembers were asked to report children from all pregnancies they had ever had or fathered. In subsequent sweeps at age 42, 46 and 50, only new children since the last interview were reported, and these were added to the total number of children since last participation. The ages of children reported previously were updated using their date of birth and the date of the current interview. It was only possible to derive the number of children cohort members had with a previous partner at age 33, as this was the only sweep where all children were reported along with information on who the other parent is. In later sweeps, when only new children since last interview are reported, information on the other parent is only available for these new children, but not for children reported in a previous sweep.

Variables for non-biological children were derived from the household grid, meaning that only those living in the household at the time of the follow up were captured.

Annotated Stata code is available here for further inspection of the variable derivation process in the NCDS.

3.4 Date errors and inconsistencies in the NCDS

Data from reported pregnancies was supplemented by information from the household grid in instances where more biological children were reported living in the household than were reported as pregnancies. Surplus children were simply added to those reported as pregnancies, and the number of boys and girls was also adjusted, as was the age of the eldest and youngest child. For each sweep, a flag variable has been created that indicates where household grid data has been used to supplement. This applied to 350 cases at age 33, 922 at age 42, 271 at age 46, and 198 at age 50. This adjustment method was not used at age 23 as it only applied to one case.

Data were checked for inconsistencies between the total number of biological children reported compared to the previous sweep. A variable has been created that flags cases where more children are reported in the previous sweeps compared to the current sweep. This applied to 22 cases at age 33, whilst there were no inconsistencies at any other ages.

4. Fertility histories in 1970 British Cohort Study (BCS70)

4.1 About the BCS70

The 1970 British Cohort Study is an ongoing birth cohort study, that has followed the lives of around 17,000 people born in England, Scotland, and Wales in a single week in April of 1970. The initial survey was carried out at the birth of the child, with follow-ups at age 5, 10, 16, 26, 30, 34,38, 42, 46, and 51 years. As a longitudinal and multidisciplinary study, a large amount of varied and rich information has been collected about cohort members through their lives. The BCS70 has contributed significantly to research across the scientific community and to key policy areas. Further information about BCS70 can be found on the website for the Centre for Longitudinal Study (CLS).

4.2 Original data that were used from the BCS70

The fertility dataset is derived from data covering ages 26, 30, 34, 38, 42, 46, and 51 (sweeps 5 to 11), which were collected between 1996 and 2021. These original data are available under End User Licence from the UK Data Service.

Citation for the BCS70 datasets is:

University College London, UCL Social Research Institute, Centre for Longitudinal Studies. (2024). *1970 British Cohort Study*. [data series]. *11th Release.* UK Data Service. SN: 200001, DOI: http://doi.org/10.5255/UKDA-Series-200001

Table 4.1: Overview of original data used for derivation of fertility variables at each sweep in the BCS70

Sweep	Data sections used for derivation [data file name]	Sample size
Age 26 (Sweep 5, 1996) Survey mode: Postal	-Main questionnaire *[bcs96x]:* any children and number, marital status and cohabitation. -Household grid *[bcs96x]:* sex, age, and relationship to CM of all persons living in the household at sweep.	9,003[a]
Age 30 (Sweep 6, 2000) Survey mode: Face to face	-Main questionnaire *[bcs2000]:* marital status and cohabitation, interview date. -Pregnancy data *[bcs2000]:* all pregnancies ever, outcome of pregnancy, sex, month and year of birth, where child is now, other parent of child. -Household *[bcs2000]:* sex, age, and relationship to CM of all persons living in the household at sweep.	11,261
Age 34 (Sweep 7, 2004) Survey mode: Face to face	-Main questionnaire *[bcs_2004_followup]:* marital status and cohabitation, interview date. -Pregnancy data *[bcs_2004_followup]:* new pregnancies since last interview, outcome of pregnancy, sex, month and year of birth, where child is now, other parent of child. - Absent child grid *[bcs_2004_followup]*: biological children who no longer live in the household, including information on the other parent. -Household grid *[bcs_2004_followup]:* sex, age, and relationship to CM of all persons living in the household at sweep.	9,665
Age 38 (Sweep 8, 2008) Survey mode: Telephone	-Main questionnaire *[bcs_2008_followup]:* marital status and cohabitation, interview date. -Pregnancy data *[bcs_2008_followup]:* new pregnancies since last interview, outcome of pregnancy, sex, month and year of birth, where child is now, other parent of child. -Household grid *[bcs_2008_followup]:* sex, age, and relationship to CM of all persons living in the household at sweep.	8,874

Sweep	Data sections used for derivation [data file name]	Sample size
	-Absent child grid *[bcs_2008_followup]:* children who no longer or never lived in the household, including information on the other parent.	
Age 42, (Sweep 9, 2012) Survey mode: Face to face	-Main questionnaire *[bcs70_2012_flatfile]:* interview date -Derived variables *[bcs70_2012_derived]:* marital status and cohabitation. -Person grid *[bcs70_2012_persongrid]:* any person ever reported living with CM in current or previous sweeps, and any absent children or new children (sex, month and year of birth, relationship to CM, whether lives in household, whether child is current partner's child).	9,841
Age 46 (Sweep 10, 2016) Survey mode: Face to face	-Main questionnaire *[bcs_age46_main]:* interview date. -Derived variables *[bcs_age46_main]:* marital status and cohabitation. -Person grid *[bcs_age46_persongrid]:* data on any person ever reported living with CM in current or previous sweeps, and any absent children or new children (sex, month and year of birth, relationship to CM, whether lives in household, whether child is current partner's child).	8,581
Age 51 (Sweep 11, 2021) Survey mode: Face to face	-Main questionnaire *[bcs11_age51_main]* : interview date, marital status and cohabitation. -Person grid *[bcs11_age51_persongrid_longf]:* data on any person ever reported living with CM in current or previous sweeps, and any absent children or new children (sex, month and year of birth, relationship to CM, whether lives in household, whether child is current partner's child).	8,016

Notes: [a] The age 26 interview was a postal questionnaire and the first time the cohort members had been contacted since age 16, so a smaller sample than what was achieved later at age 30. The smaller sample responding at age 26 was overrepresented in terms of those having a university degree qualification, and under representative of those with no qualifications, compared to the sample at age 30.

The full guides to the datasets for each BCS70 sweep are available from the CLS website via the following weblinks:

Sweep 5, Age 26, 1996

Sweep 6, Age 30, 2000

Sweep 7, Age 34, 2004

Sweep 8, Age 38, 2008

Sweep 9, Age 42, 2012

Sweep 10, Age 46, 2016

Sweep 11, Age 51, 2021

4.3 How fertility variables were derived in the BCS70

Variables on biological children at age 26 to 38 were mainly derived from pregnancy data where cohort members reported outcomes of conceived or fathered pregnancies. For each live birth, information was collected on the sex of the child and who the other parent is. At age 26 the information was less detailed with information only on whether cohort members had had any children and the number, so the household grid was used to provide information on children's ages and sex, meaning that any biological children not living in the household would not be included in these two variables. At age 30 the information was more comprehensive, with a full report on any children ever had (only live births included in derivation), including information on their sex, month and year of birth, whether the child lived in household, elsewhere, or was no longer alive; and whether current partner is the other parent. From age 34 and onwards, only pregnancies since the last interview were reported, providing similar information on each child as at age 30. These new children were added to the total number of children counted since the last survey participation. The ages of children reported previously were updated using their date of birth and the date of the current interview.

At age 42, 46 and 51 the main data use for derivation was the person grid. This contains all persons the cohort member has lived with in the current or previous

sweeps, and any absent children or new children since participation in the last survey. For each child in the grid, information was available on their relationship to the cohort member, year and month of birth, sex, whether they lived in the household, and their relationship to the cohort member's current partner in the household.

Variables on non-biological children were derived from the household grid at age 26 to 38, meaning that only those living in the household at the time of the follow up were captured. At age 42, 46 and 51, the person grid was used for variables on non-biological children, and for consistency with the previous sweeps, only the non-biological children who currently lived in the household were included.

Annotated Stata code is available here for further inspection of the variable derivation process in the BCS70.

4.4 Data errors and inconsistencies in the BCS70

At age 30, 34 and 38, the household grid was used to supplement cases where more biological children were reported living in the household than reported as pregnancies. These surplus children were added to those reported as pregnancies, the number of boys and girls was adjusted, and the age of the variables for the eldest and youngest child also incorporated the surplus household children. A variable has been created that flags the cases where the household grid provides additional information. This applied to 80 cases at age 30, 384 cases at age 34, and 117 cases at age 38. Because the person grids were used to derive all fertility variables at age 42, 46 and 51 such inconsistencies could not be captured. At age 26 this method was not used as fewer people had completed the household grid than had reported their number of biological children.

At age 30, 34, 38, 42, 46 and 51, data were checked for inconsistencies between the total number of biological children reported compared to the previous sweep. A variable has been created that flags cases where more biological children are reported in the previous sweeps compared to the current sweep. This applied to 44 cases at age 30, 149 cases at age 42, 47 cases at age 46, and 68 cases at age 51, whilst there were no inconsistencies at age 34 and 38.

4.5 Survey and non-response weights in BCS70

The BCS70 fertility dataset includes a weighting variable designed to correct for non-response in the age 51 survey. Note that similar weights are not available for the other sweeps. This weight should be used in any analyses of these data, with further information provided in the data user guide at age 51 (a link is provided further above).

Example Stata code of application of weights at age 51

```
prop anybiochildren_51 [pweight= bd11weight_main]
```

5. Fertility histories in Next Steps

5.1 About Next Steps

Next Steps was previously known as the Longitudinal Study of Young People in England (LSYPE). It began following around 16,000 young people who were in Year 9 in 2004 in state or independent schools across England, born mainly in 1989 and 1990. Since the initial survey when cohort members were around age 14, there have been annual follow up surveys until age 20, and after this at age 25 and 32. The initial focus of the study was on cohort members' schooling and their further transition into education and employment. Increasingly it has taken a multidisciplinary approach with inclusion of a wide range of other aspects, including physical and emotional health, wellbeing, social participation, attitudes, and family life. Further information about Next Steps can be found on the website for the Centre for Longitudinal Study (CLS).

5.2 Original data that were used from Next Steps

The fertility dataset is derived from data covering age 26 and 32 (sweeps 8 and 9), which were collected in 2015 and 2022. These original data are available under End User Licence from the UK Data Service.

Citation for the Next Steps datasets is:

University College London, UCL Social Research Institute, Centre for Longitudinal Studies. (2024). *Next Steps (also known as the Longitudinal Study of Young People in England)*. [data series]. *12th Release.* UK Data Service. SN: 2000030, DOI: http://doi.org/10.5255/UKDA-Series-2000030

Table 5.1: Overview of original data used for derivation of fertility variables at each sweep in Next Steps

Sweep	Data sections used for derivation [*data file name*]	Sample size
Age 25 (Sweep 8, 2015) Survey mode: Mixed mode (web, telephone and face to face)	-Main questionnaire *[ns8_2015_main_interview]*: marital status and cohabitation, interview date. - Child grid *[ns8_2015_children]*: month and year of birth, sex, relationship to CM, whether in HH, whether cohabiting partner's child, all children CM considers themselves a parent to. -Household grid *[ns8_2015_household_members]:* sex, age and relationship to CM of all persons living in the household at the interview (with exclusion of children already reported in the child grid). - Main data sweep 1 [*wave_one_lsype_young_person_2020*]: Sex and months and year of birth of CM. - Longitudinal file *[ns9_2022_longitudinal_file]*: weights	7,707
Age 32 (Sweep 9, 2022) Survey mode: Online, face-to-face, video, telephone	-Main questionnaire *[ns9_2022_main_interview]:* interview date -Derived variables *[ns9_2022_derived_variables]:* Marital status and cohabitation. - Person grid *[ns9_2022_person_grid]:* Details of all persons currently or previously living in the household as well as all children reported in current or previous sweep (sex, month and year of birth, relationship to CM, and whether or not currently in the household). -Children with non-resident parent *[ns9_2022_children_with_non_resident_parent]:* Data on all the children with a non-resident parent, a grid ID number links these to the person grid. - Main data sweep 1 *[wave_one_lsype_young_person_2020]:* Sex and months and year of birth of CM. - Longitudinal file *[ns9_2022_longitudinal_file]*: survey weights	7,279

The guides to the datasets for each of the Next Steps sweeps are available from the CLS website via the following weblinks:

Sweep 8, Age 25, 2015

Sweep 9, Age 32, 2022

5.3 How fertility variables were derived in Next Steps

At age 25, variables on biological children were derived from the child grid, which are all the children that the cohort members consider themselves to be a parent to. For each child, there is information on their sex, month and year or birth, whether the child lives in household, their relationship to the cohort member, and whether the current cohabiting partner is the other parent. Variables on non-biological children at age 25 were derived from the child grid (for those non-biological children the CM considered themselves to be a parent of); with the addition of children from the household grid (other children not reported in the child grid but currently living with the CM), which contained details on their sex, year and month of birth, and their relationship to the cohort member. Only non-biological children who lived in the household were included to be consistent with the derived variables in the other cohorts.

At age 32, variables on biological and non-biological children were derived largely from the person grid, which had information on all persons currently or previously living in the household and all children reported in any of the sweeps. Information includes the child's sex, month and year of birth, relationship to CM, and whether currently in the household. For biological children, data were linked to a dataset on children with a non-resident parent, to derive the variables on children by previous partners. For non-biological children, again only those residing in the household at age 32 were included in the derived variables.

Annotated Stata code is available here for further inspection of the variable derivation process in Next Steps.

5.4 Data errors and inconsistencies in Next Steps

The data were checked for any inconsistency between the total number of children reported at age 32 compared to at age 25. A variable has been created that flags cases where more biological children are reported at age 25 than at age 32. This applies to 75 cases. Any inconsistencies between the number of biological children reported as in the child grid at age 25 and the household grid at age 25 could not be established because the household grid excludes children already reported in the child grid. Similarly at age 32, the person grid used for derivation of fertility variables combines all reported biological children ever and all persons that had ever lived in the household, so any inconsistencies between number of biological children reported as pregnancies and those living in the household could not be identified.

5.5 Survey and non-response weights in Next Steps

The Next Steps fertility dataset includes variables that reflect the complex sampling design of the initial survey as well as non-response weights for sweeps at age 25 and 32. These should be used in any analyses of these data, with further information provided in the data user guides on the use of weights (links are provided further above).

Example Stata code of application of weights at age 25

```
replace W8FINWT=. if W8FINWT<0  //non-applicable weights coded as missing
svyset SAMPPSU [pweight= W8FINWT], strata(SAMPSTRATUM)
svy: proportion anybiochildren_25
```

Example Stata code of application of weights at age 32

```
replace W9FINWT=. if W9FINWT<0  //non-applicable weights coded as missing
svyset SAMPPSU [pweight= W9FINWT], strata(SAMPSTRATUM)
svy: proportion anybiochildren_32
```

6. Description of the harmonised datasets

6.1 Licensing and data access

Datasets on fertility histories in the NSHD, NCDS, BCS70, and Next Steps have been processed by CLS and supplied to the UK Data Service. All data users need to be registered with the UK Data Service and to sign the UKDS End User Licence before they can download the data. Details of how to do this are available at ukdataservice.ac.uk/get-data/how-to-access/registration.

The NCDS, BCS70 and Next Steps fertility datasets are available as safeguarded data, which can be downloaded from the UK Data Service once the End User Licence (EUL) access conditions have been accepted by the user.

The NSHD fertility data can be accessed by downloading the UKDS Special Licence application form. Once the form has been reviewed by UKDS and approved by the NSHD Data Sharing Committee the data will be available to download. For accessing and linking to other NSHD data see section 6.2.

Access for additional NSHD data

The NSHD fertility histories dataset is also available from MRC Unit for Lifelong Health and Ageing at UCL (LHA), which manages the NSHD. This route of access is necessary for analysts wishing to use the fertility data alongside other information held for the 1946 cohort. The research project needs to first be approved by the NSHD Data Sharing Committee. Full details on how to access the data can be found here. Once a data access form has been approved and a data sharing agreement is in place, the data can be accessed via www.condor.ucl.ac.uk .

6.2 List of datasets

Datasets are all in wide/flat format. All datasets are listed in Table 6.1 below.

Table 6.1: List of available datasets

Name of the dataset	Content summary
harmonised_fertility_histories_nshd	The dataset contains fertility variables derived in the NSHD at age 19, 20, 22, 26, 31, 36, 43 and 53.
harmonised_fertility_histories_ncds	The dataset contains fertility variables derived in the NCDS at age 23, 33, 42, 46 and 50.
harmonised_fertility_histories_bcs	The dataset contains fertility variables derived in the BCS70 at age 26, 30, 38, 42, 46 and 51.
harmonised_fertility_histories_nextsteps	The dataset contains fertility variables derived in Next Steps at age 25 and 32.

For the NCDS, BCS70, and Next Steps, other data from these cohorts are also available via the UKDS. 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.

6.3 Harmonised variables syntax

The code developed to derive all harmonised fertility variables is available on the All code files can be found on the CLS Data GitHub page at https://github.com/CLS-Data/Fertility-histories-in-four-UK-cohort-studies

6.4 Overview of the fertility variables

Table 6.2 provides an overview of the fertility variables for each cohort at each sweep. This enables data users to see where information overlaps across cohorts and where there are gaps.

Table 6.2: Fertility variables for each cohort and their sweeps

Variable name	Variable label	NSHD								NCDS					BCS70							NS	
		Age 19	Age 20	Age 22	Age 26	Age 31	Age 36	Age 43	Age 53	Age 23	Age 33	Age 42	Age 46	Age 50	Age 26	Age 30	Age 34	Age 38	Age 42	Age 46	Age 51	Age 25	Age 32
sex*	sex of cohort member									✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
cmbyear*	Birth year of CM									✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
cmbmonth*	Birth month of CM									✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
survey		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
intyear	Interview year				✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓
intmonth	Interview month				✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓
partner	Whether has a partner in HH									✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
marital	Marital status	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
anybiochildren	Whether has had any bio children	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
biochild_tot	Number of bio children	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
biototal_flag	More biological children reported at previous age than at current age		✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓		✓	✓	✓	✓	✓	✓		✓
biochildhh_total**	Number of bio children in HH						(✓)	(✓)		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
biochild_extra_flag	Flag: More bio children reported in HH grid than in pregnancy data										✓	✓	✓	✓		✓	✓	✓					
biochildnonhh_total	Number of bio children not in HH									✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
biochildprev_total	Number of bio children had with a previous partner										✓					✓	✓	✓	✓	✓	✓	✓	✓

Variable name	Variable label	NSHD								NCDS					BCS70							NS	
biochildprevany	Have had any bio children with a previous partner									✓				✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
biochildy_eldest	Age in years of eldest bio child				✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
biochildy_youngest	Age in years of youngest bio child				✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
cmageybirth_eldest	Age in years of CM at birth of eldest bio child				✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
cmageybirth_youngest	Age in years of CM at birth of youngest bio child				✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
biochildboy_total	Number of bio children who are boys				✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
biochildgirl_total	Number of bio children who are girls				✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
anynonbio	Whether has any non-bio children in HH									✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
nonbiochild_tot	Number of non-bio children in HH									✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
adopt_tot	Number of adopted children in HH										✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
foster_tot	Number of fostered children in HH									✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
step_tot	Number of stepchildren in HH									✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
nonbiochildy_eldest	Age in years of eldest non-bio child									✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
nonbiochildy_youngest	Age in years of youngest non-bio child									✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
nonbiochildboy_total	Number of non-bio children who are boys									✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
nonbiochildgirl_total	Number of non-bio children who are girls									✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
anychildren	Whether has any children (bio or non-bio)						✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
children_tot	Number of children (bio or non-bio)						✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
childy_eldest	Age in years of eldest child (bio or non-bio)									✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Variable name	Variable label	NSHD								NCDS					BCS70							NS	
childy_youngest	Age in years of youngest child (bio or non-bio)									✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
childboy_total	Number of children who are boys (bio or non-bio)									✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
childgirl_total	Number of children who are girls (bio or non-bio)									✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
partnerchildbio	Whether has live-in partner/spouse and/or any bio children	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
partnerchildany	Whether has live-in partner/spouse and/or any bio or non-bio children									✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
cflag	Mismatched information in anybiochildren and biochild_tot		✓				✓	✓															
cflag_19	Biological children reported at previous age(s), not at current age		✓	✓	✓	✓																	
cnflag	Fewer biological children reported at current age than previous age(s)				✓	✓	✓	✓															
cgflag	Mismatched information in anybiochildren and biochildgirl_total				✓	✓	✓																
cbflag	Mismatched information in anybiochildren and biochildboy_total				✓	✓	✓																

Notes: In the datasets, variable names have a suffix that identifies the sweep. * Time invariant variable, included once in each dataset rather than for each age sweep. ** In the NSHD, total number of biological children in household (biochildhh_total) cannot be separated from non-biological children.

6.5 Identifiers

Individual identifiers

The individual identifier variables in each of the cohorts are NSHDID_UKDS01, NCDSID, BCSID, NSID.

Use of individual identifiers to merge with cohort study data

For NCDS, BCS70, and Next Steps, the data are identified with the same research IDs used for the rest of cohort data available at the UKDS. This enables the derived dataset to be easily merged with additional variables.

While merging Covid-19 survey data with cohort study data should be similarly straightforward using their respective identifiers, users should consult individual user guides for specific information beforehand.

For NSHD, each researcher [or team of researchers working on the same project] has an unique serial number generated for their use. This serial number is attached to the data they initially request from Condor and for any subsequent data they request so that all NSHD datasets can easily be merged together.

6.6 Missing values

For missing values, the following coding has been applied: -100 no participation in sweep, -99 information not provided.

6.7 Data errors and inconsistencies

Users should be aware of the following data errors and inconsistency details.

For some cases there were more biological children reported as living in the household than those reported as pregnancies. These additional biological children were then added to the total number of biological children reported as pregnancies. A variable (biochild_extra_flag) has been created which flags the cases where this method was used. Checks were also made between sweeps in term of the number of biological children reported. For some cases more biological children were

reported at the previous sweep than at the current sweep. A variable (biototal_flag) has been created which flags these cases.

The extent of these inconsistencies have been outlined previously for each of the cohorts.

For the NSHD data, additional flag variables have been derived to indicate if there was a mismatch in information recorded in different variables collected at the same sweep of data collection, e.g. if the study member reports to have no child in a variable indicating if they have a child or not, but reports to have one or more children in a variable indicating the total number of children they have.