Centre for Longitudinal Studies

CLS Cohort Studies

Data Note 1

Longitudinal Linkage in BCS70:
Rationalising Case Identifiers

Brian Dodgeon

Centre for Longitudinal Studies
Institute of Education
20 Bedford Way
London WC1H 0AL
Tel: 020 7612 6860
Fax: 020 7612 6880
Email cls@cls.ioe.ac.uk
Web http://www.cls.ioe.ac.uk

CLS COHORT STUDIES

DATA NOTE 1

Longitudinal Linkage in BCS70: Rationalising Case Identifiers

Brian Dodgeon

December 2002

Introduction

This document has been prepared to accompany a deposit, with the UK Data Archive at the University of Essex, of data from the 3 most recent follow-ups (age 16, 26 and 30) of the 1970 British Cohort Study (BCS70), which is a continuing, multidisciplinary, national, longitudinal study.

BCS70 began when data were collected about the births of families of 17,198 babies born in England, Scotland, Wales and Northern Ireland in the week 5-11 April 1970. Since the birth survey there have been five other major data collection exercises in order to monitor their health, education, social and economic circumstances. These were carried out in 1975 (age 5), 1980 (age 10), 1986 (age 16), 1996 (age 26), and 1999/2000 (age 30). Sub-samples have also been studied at various ages: for example at age 21, a 10 per cent representative sample was assessed for basic skills difficulties.

From its original focus on the circumstances and outcomes of birth, BCS70 has broadened in scope to map all aspects of health, education and social development of their subjects as they passed through childhood and adolescence. In later sweeps, the information collected has covered their transitions into adult life, including leaving full-time education, entering the labour market, setting up independent homes, forming partnerships and becoming parents.

The latest round of data collection for BCS70 took place in 1999/2000 when cohort members were aged 29/30. The main aim of these most recent surveys was to explore the factors central to the formation and maintenance of adult identity in each of the following domains:

- Lifelong learning
- Relationships, parenting and housing
- Employment and income
- Health and health behaviour
- Citizenship and values

Cohort Studies User Support Group

This provides advice and guidance on the use of Cohort Studies data; produces documentation; collates and disseminates information on uses of the data, publications, and other developments; produces and distributes a newsletter and working papers; provides access to non-computerised Cohort Studies data; and collects additional information.

Contacting the User Support Group

The User Support Group can be contacted by post, 'phone, fax, or email as shown below:

Cohort Studies User Support Group,
Centre for Longitidinal Studies,
6th Floor: Institute of Education,
20 Bedford Way,
London WC1H 0AL

Tel: +44 0207 612 6864
Fax: +44 0207 612 6880
Email: cohort@cls.ioe.ac.uk
Internet: http://www.cls.ioe.ac.uk/Cohort/Ncds/mainncds.htm

Acknowledgements

I am indebted to Andrew Cullis, my colleague at CLS, who began the work on rationalising the case identifiers in the early months of this year, and assembled a good deal of useful information, from which I was able to proceed after Andrew left CLS to take up a post with the Scottish Longitudinal Study.

Background to Longitudinal Linkage Problem in BCS70

The difficulty in linking the various sweeps longitudinally has arisen in the context of a dual numbering system, which was adopted for historic reasons connected with the ownership of the data. Originally the cases were referenced by ***Key number*** (which could be anything from 2 to 6 digits), but a parallel ***BCS70 serial number*** (8 digits) was introduced in later sweeps.

The essence of the problem was that between 30 and 40 cases had a BCS70 serial number but no apparent Key number, so there was no reliable way for researchers to know if they had been interviewed at earlier sweeps. There were also a small number of anomalous cases where one Key number seemed to correspond to two distinct BCS70 serial numbers, or vice versa.

An exercise was undertaken at the Centre for Longitudinal Studies during 2002 to investigate all these cases, with the aim of re-instituting ***Key number*** as the comprehensive unique identifier for longitudinal linkage. This document explains the methodology, and the appendices list the cases which have been tidied up, and the SPSS syntax used.

Identification by KEY number

The data held at the UK Data Archive for the British Cohort Study 1970 are in the form of the 6 separate main BCS70 sweeps: at ages 0, 5, 10, 16, 26 and 30 (there were also two 10% sample sweeps at ages 22 months and 36 months, whose data are amalgamated with the birth (age 0) data. Other small sample surveys have not yet been fully documented for deposit).

Longitudinal linkage of the first four sweeps has always been possible with little difficulty, because all members surveyed, without exception, have a unique 6-digit identifier in the variable **KEY**.

In the 16-year-old dataset, as originally deposited, the key number as such was not present, but could be derived by combining the 5-digit **CHESNO** variable with the 1-digit **TC2** (twin code) variable. In the revised deposit of the data (Dec 2002), this process has now been done, so that a variable KEY exists at age 16.

Those members present at the original 1970 survey have KEY numbers from 10 to 206210, and include 626 children living in Northern Ireland. After this initial survey, the Northern Ireland population was excluded from all subsequent sweeps, except for the small minority who had moved to Great Britain in the meantime.

Attempts continued for sixteen years to locate and contact children born in the week 5-11 April 1970 who may have been missed by the birth survey, including those born abroad who subsequently moved to Britain. As a result, additional members were added to the survey at the 5-year-old sweep (KEY nos. 300010 to 450490), the 10-year-old sweep (KEY nos. 600020 to 703560), and the 16-year-old sweep (KEY nos. 800020 to 804890) [See Appendix 7]. This numbering system does not always exactly reflect the first appearance of a cohort member's data chronologically, as the key number will have been allocated when the member was located, but they may not have been successfully interviewed till a later sweep.

After the 16-year-old sweep, this process of attempting to expand the population base was limited simply to going back to those already located but not successfully interviewed.

Identification by BCS70 serial number

In the late 1980s, a parallel identifier was introduced, by concatenating a 5-digit ‘Y’ number with the 1-digit twin code and a 2-digit ‘check digit’ to form the 8-digit **BCS70 serial number** (variable SERIAL in the 26-year dataset, BSERIAL at 30 years).

SERIAL appears for the first time as a variable in the 26-year dataset, with every case uniquely coded. KEY was also present, but 27 cases appeared with KEY=0 (see Appendix 1). Four of these were found to have had their serial numbers keyed incorrectly, of which two could then be linked to a KEY number by reference to the corresponding serial number in the 30-year-old dataset, leaving a net total of 25 (see Appendix 1).

In the 30-year-old dataset, all cases have a unique value of BSERIAL, but 35 cases appeared with KEY coded as system-missing (see Appendix 2). 14 of these were the same cases that had KEY=0 at the 26-year survey, and a further 8 could be indirectly linked to a KEY number by matching with the corresponding SERIAL number at the 26-year-old data (see Appendix 2); so the net figure of additional problematic cases at the 30-year-old sweep was 13 (i.e. 35 - 14 - 8).

Combining these 13 with the 25 problematic cases from the 26-year-old data, there were 38 cases which had no valid value for KEY.

Matching through Address Database

The above figure of 38 was reduced further, by reference to the internal confidential address database system held at the Centre for Longitudinal Studies, where 8 key numbers were found which had not previously been cross-referenced with a serial number in the data (see Appendix 3).

This left a hard core of **30** cases which could still not be linked to a key number (see Appendix 5).

Cleaning of 2 other spurious cases

In examining the above, two instances were identified of cases where one KEY number seemed to correspond to more than one serial number (see Appendix 4.).

In one case they turned out to be the same person, so the 26-year serial number was changed to harmonise with the 30-year serial.

In the other case, it was two different cohort members: the Key number of one had been mis-typed at the 30-year survey, so this has now been corrected (see Appendix 4).

[These two cases were not contained in the 30 remaining from the above process].

Investigation of origin of remaining 30 cases with no KEY number

An extensive search was conducted through a variety of media at CLS to determine whether the 30 outstanding cases with no match to a key number might possibly be returning members who had been interviewed at earlier sweeps, possibly under a different name.

(a) Attempt to match on name/NHS number(old and new)/phone number using confidential address database

No other members on the address database could be found whose details matched these 30 cases in any of the above respects.

However, the address database was set up in the 1980s from the records of people who were still in touch at that time: it contains a total of 16,764 records, compared with 18,733 in the combined dataset obtained by linking all data longitudinally. So there are around 2,000 members who have participated at some point, but who are not on the address database. This left open the possibility that the 30 cases may be cohort members who dropped out in the early years of the survey, but recently came back.

(b) Reference to card index systems

There are three historical card index sets housed at the CLS: one ordered by key number, and the other two ordered by surname at birth and at 16. None of the 30 cases could be found in either of the two surname-indexed card systems.

(c) Scrutiny of the paper questionnaires stored in the IoE from the 26-year survey.

An attempt was also made to find the paper questionnaires of the 23 (of the 30) cases who were known to have been interviewed at the 26-year survey, to see if there was any further information which might link them to earlier sweeps; but currently the physical storage configuration is not conducive to locating cases systematically on serial number or name, so it was not possible to find them at the present time.

(d) Consideration of the possibility they might be Northern Ireland cases.

As noted in the section 'Identification by Key number' 626 children were interviewed in Northern Ireland at the birth survey, but never again (unless they subsequently moved to Britain).

There was a possibility that these 30 cases might be people who subsequently moved to Britain from Northern Ireland. But scrutinising the names, this seems very unlikely, as 13 of the 30 have names indicative of ethnic minorities, which would be somewhat inconsistent with the NI demography.

Another factor is that only 21 original Northern Irelanders with KEY numbers have been found to have migrated to Britain during the last thirty years (see Appendix 10), so it seems unlikely that as many as 30 others turned up around the time of the last two surveys.

Conclusion

Having exhausted all the above avenues, it was impossible to escape the conclusion that these 30 cases are simply additional cohort members, who had not been interviewed until the 26- or 30-year-old surveys.

As such, there is no possibility of longitudinal linkage to the first four main sweeps, but in this respect they are no different from the 97 other cohort members with valid Key numbers whom we didn't succeed in contacting until 26 or later [see Appendix 7. (97 = 79 + 48 – 30)].

A policy decision was therefore taken to leave these cases in the data and to allocate new Key numbers to them. Although at present they do not yield as much longitudinal information as the vast majority of cohort members, hopefully over time they will contribute data in many more sweeps. Researchers always have the option of leaving them out of analyses if they wish.

They were given the consecutive Key numbers 900010 to 900300 (see Appendix 8).

As a result of this cleaning process, all 18,733 BCS70 members who have ever been interviewed now have unique Key numbers, and all duplication has been eliminated. We therefore recommend researchers ignore the BCS70 serial number from now on and use ***Key*** as the unique identifier for longitudinal linkage.

The SPSS syntax to clean the 26-year and 30-year data is listed in Appendix 8. These were the only two sweeps which needed to be altered, but for the sake of completeness, the 16-year data has also been re-deposited with the addition of a specific ***Key*** variable, so that researchers will no longer have to combine the 5-digit CHESNO and 1-digit TC2 variables to arrive at the 6-digit identifier.

Appendix 1

BCS70 26-year-old dataset.

26-year serial numbers (SERIAL) of those cases which had KEY coded as 0.

00433049* 00617905** 02039000# 16032893## 21033000 21056000
21178000 21225000 21291000 21320000 21348000 21369000
21391000 21397000 21406000 21413000 21530000 21539000
21575000 21583000 21600000 21623000 21626000 21658000
21667000 21767000 21856000

27 cases in total (there are 9 other cases which have no KEY number as such in the data, but do have the 5-digit CHESNO and 1-digit CTC code, from which KEY can be derived. These have been tidied up with a proper key number in the re-deposited data).

* this serial number was in fact mis-typed into the 26-year-old data as originally deposited. The number should have been 00933049, and has now been amended. It is matched to KEY number 129360 in the 30-year-old dataset.

** this serial number was also mis-typed: it should have been 06179057. This has now been amended.

# this serial number was mis-typed: it should have been 20339000. This has now been amended. It is matched to KEY number 102260 in the 30-year-old dataset.

## this serial number was mis-typed: it should have been 13032093. This has now been amended.

Appendix 2

BCS70 30-year-old dataset.

30-year serial numbers (BSERIAL) of those cases which had KEY coded as system-missing.

21033000 21051000 21055000 21056000 21059000 21178000
21179000 21185000 21225000 21239000 21291000 21320000
21338000 21369000 21387000 21391000 21401000 21413000
21421000 21453000 21530000 21556000 21581000 21600000
21619000 21620000 21623000 21644000 21655000 21667000
21673000 21689000 21740000 21753000 21856000

35 cases in total.

Of these 35, the following 14 serial numbers coincided with 26-year serial numbers with KEY=0 (see Appendix 1):

21033000 21056000 21178000 21225000 21291000 21320000
21369000 21391000 21413000 21530000 21600000 21623000
21667000 21856000

A further 8 cases, although having no cross-reference to KEY number in the 30-year-old dataset, corresponded with serial numbers in the 26-year-old dataset which *were* cross-referenced to KEY:

21051000 21179000 21185000 21239000 21620000 21655000
21673000 21689000

Appendix 3

KEY numbers found by reference to the CLS confidential address database system

Serial no.	**Key no.**
06179057	144810
13032093	7830
21421000	301340
21556000	24550
21619000	83040
21644000	610230
21740000	150150
21753000	608780

8 cases in total.

Appendix 4

Two cases where one KEY number corresponded to more than one SERIAL (or BSERIAL) number.

Case (a)

KEY	SERIAL (26-year survey)	BSERIAL (30-year survey)
83250	20458000	00128031

On examining these two entries on the confidential address database, they turned out to be the same person. Although the surnames listed for each entry were different, both her two forenames were identical, as was the day of birth.

26-year serial no. 20458000 was therefore **altered** to 00128031

Case (b)

KEY	SERIAL (26-year survey)	BSERIAL (30-year survey)
80430		08266089 & 18074999

In contrast with case (a), this turned out to be two different cohort members, where a KEY number which should have been **804830,** had been erroneously typed into the data as **80430.**

The KEY number corresponding to 30-year serial number 18074999 was therefore **altered** to 804830.

Appendix 5

Serial numbers of the 30 cases at the 26-year or 30-year survey for which no KEY number could be found.

Present at 26-year survey only:

21348000 21397000 21406000 21539000 21575000 21583000
21626000 21658000 21767000

9 cases in total.

Present at both 26-year and 30-year surveys:

21033000 21056000 21178000 21225000 21291000 21320000
21369000 21391000 21413000 21530000 21600000 21623000
21667000 21856000

14 cases in total.

Present at 30-year survey only:

21055000 21059000 21338000 21387000 21401000 21453000
21581000

7 cases in total.

Appendix 6

BCS70 members in Northern Ireland at the time of the birth survey, who appeared later in Britain.

	Present	Not Present	Total
Birth survey	626	0	626
5-year-old survey	11	615	626
10-year-old survey	18	608	626
16-year-old survey	15	611	626
26-year-old survey	10	616	626
30-year-old survey	9	617	626
Appeared at any survey post-1970	21	605	626

Appendix 7

Cumulative total of BCS70 members interviewed, showing additions to survey at each successive time-point (*after* this cleaning process)

	Cumulative Total Cases
Birth survey (incl.22-month & 32-month follow-ups):	**17,196**
5-year-old survey contains 292 cases not previously seen:	**17,588**
10-year-old survey contains an additional 847 cases:	**18,435**
16-year-old survey contains an additional 171 cases:	**18,606**
26-year-old survey contains an additional 79 cases:	**18,685**
30-year-old survey contains an additional 48 cases:	**18,733**

Appendix 8

SPSS Syntax to rationalise the Case Identifiers in the last three BCS70 data sweeps

30-year dataset:

```
If (bserial='21033000')key=900010.
If (bserial='21055000')key=900020.
If (bserial='21056000')key=900030.
If (bserial='21059000')key=900040.
If (bserial='21178000')key=900050.
If (bserial='21225000')key=900060.
If (bserial='21291000')key=900070.
If (bserial='21320000')key=900080.
If (bserial='21338000')key=900090.
If (bserial='21348000')key=900100.
If (bserial='21369000')key=900110.
If (bserial='21387000')key=900120.
If (bserial='21391000')key=900130.
If (bserial='21397000')key=900140.
If (bserial='21401000')key=900150.
If (bserial='21406000')key=900160.
If (bserial='21413000')key=900170.
If (bserial='21453000')key=900180.
If (bserial='21530000')key=900190.
If (bserial='21539000')key=900200.
If (bserial='21575000')key=900210.
If (bserial='21581000')key=900220.
If (bserial='21583000')key=900230.
If (bserial='21600000')key=900240.
If (bserial='21623000')key=900250.
If (bserial='21626000')key=900260.
If (bserial='21658000')key=900270.
If (bserial='21667000')key=900280.
If (bserial='21767000')key=900290.
If (bserial='21856000')key=900300.

If (bserial='21051000')key=600280.
If (bserial='21179000')key=601630.
If (bserial='21185000')key=108930.
If (bserial='21239000')key=  76630.
If (bserial='21421000')key=301340.
If (bserial='21556000')key=  24550.
If (bserial='21619000')key=  83040.
If (bserial='21620000')key=163880.
If (bserial='21644000')key=610230.
If (bserial='21655000')key=610200.
If (bserial='21673000')key=604640.
If (bserial='21689000')key=604850.
If (bserial='21740000')key=150150.
If (bserial='21753000')key=608780.


If (bserial='18074999')key=804830.

execute.
```

26-year dataset:

```
If (serial='21033000')key=900010.
If (serial='21055000')key=900020.
If (serial='21056000')key=900030.
If (serial='21059000')key=900040.
If (serial='21178000')key=900050.
If (serial='21225000')key=900060.
If (serial='21291000')key=900070.
If (serial='21320000')key=900080.
If (serial='21338000')key=900090.
If (serial='21348000')key=900100.
If (serial='21369000')key=900110.
If (serial='21387000')key=900120.
If (serial='21391000')key=900130.
If (serial='21397000')key=900140.
If (serial='21401000')key=900150.
If (serial='21406000')key=900160.
If (serial='21413000')key=900170.
If (serial='21453000')key=900180.
If (serial='21530000')key=900190.
If (serial='21539000')key=900200.
If (serial='21575000')key=900210.
If (serial='21581000')key=900220.
If (serial='21583000')key=900230.
If (serial='21600000')key=900240.
If (serial='21623000')key=900250.
If (serial='21626000')key=900260.
If (serial='21658000')key=900270.
If (serial='21667000')key=900280.
If (serial='21767000')key=900290.
If (serial='21856000')key=900300.

If (serial='21051000')key=600280.
If (serial='21179000')key=601630.
If (serial='21185000')key=108930.
If (serial='21239000')key=  76630.
If (serial='21421000')key=301340.
If (serial='21556000')key=  24550.
If (serial='21619000')key=  83040.
If (serial='21620000')key=163880.
If (serial='21644000')key=610230.
If (serial='21655000')key=610200.
If (serial='21673000')key=604640.
If (serial='21689000')key=604850.
If (serial='21740000')key=150150.
If (serial='21753000')key=608780.

If (serial='00433049')key=129360.
If (serial='00617905')key=144810.
If (serial='02039000')key=102260.
If (serial='16032893')key=7830.

If (serial='00433049')serial='00933049'.
If (serial='00617905')serial='06179057'.
If (serial='02039000')serial='20339000'.
If (serial='16032893')serial='13032093'.
If (serial='20458000')serial='00128031'.

* The following case had the 5-digit CHESNO and 1-digit CTC defined in the data
* (from which KEY can be calculated), but for tidiness we put in the key number directly.

If (serial='21902000')key=73630.

execute.
```

16-year dataset (BCS7016):

```
compute key=10*chesno+tc2.
format key (f6.0).
format chesno (f5.0).
format sex86 (f2.0).
format lea86 (f6.0).
format dha86 (f6.0).
format regha86 (f6.0).
format land86 (f2.0).
format odoc_mt(f6.0).
format odoc_yr(f6.0).
format tdoc_mt(f6.0).
format tdoc_yr(f6.0).

variable labels key 'Unique Case Identifier'.

variable labels chesno '5-digit case identifier'.

format tc2(f2.0).

variable labels tc2 'twin code'.

value labels tc2 0  'Singleton'  1 'first of multiple birth' 2 'second of multiple birth'  3 'third of
multiple birth'.
```

16-year alpha dataset (alpha16):

```
compute key=10*chesno+tc2.
format key (f6.0).

variable labels key 'Unique Case Identifier'.

format chesno (f5.0).
variable labels chesno '5-digit case identifier'.

format tc2(f2.0).
variable labels tc2 'twin code'.
value labels tc2 0  'Singleton'  1 'first of multiple birth' 2 'second of multiple birth'  3 'third of
multiple birth'.
```

Centre for Longitudinal Studies
Bedford Group for Lifecourse and
Statistical Studies
Institute of Education
20 Bedford Way
London WC1H 0AL
Tel: 020 7612 6900
Fax: 020 7612 6880
Email cls@cls.ioe.ac.uk
Web http://www.cls.ioe.ac.uk