SLS-DSU

SCOTTISH LONGITUDINAL STUDY DEVELOPMENT & SUPPORT UNIT

Technical working paper 7

The Scottish Longitudinal Study 1936 Birth Cohort

Zengyi Huang,
Zhiqiang Feng, Chris Dibben,
Caroline Brett & Ian Deary

Published online 30 March 2017

Geography & the Lived Environment Research Institute
University of Edinburgh
Drummond Street
Edinburgh EH8 9XP

email: zengyi.huang@ed.ac.uk

Longitudinal Studies Centre Scotland
National Records of Scotland
Ladywell House
Ladywell Rd
Edinburgh EH12 7TF

email: lscs@st-andrews.ac.uk

LSCS longitudinal studies centre Scotland — LINKING LIVES THROUGH TIME

THE UNIVERSITY of EDINBURGH

University of St Andrews

National Records of Scotland

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

Author affiliations

Zengyi Huang[1,2], Zhiqiang Feng[1,2], Chris Dibben[1,2], Caroline Brett[3] & Ian Deary[3]

[1]Geography & the Lived Environmental Research Institute
University of Edinburgh
Drummond Street
Edinburgh EH8 9XP

[2]Longitudinal Studies Centre Scotland
National Records of Scotland
Ladywell House
Ladywell Road
Edinburgh EH12 7TF

[3]Centre for Cognitive Ageing and Cognitive Epidemiology
Dept of Psychology
University of Edinburgh
Edinburgh EH8 9JZ

Summary

This paper describes the creation of the Scottish Longitudinal Study (SLS) Birth Cohort of 1936 (SLSBC1936). It is structured around the existing SLS. We took the SLS birth date sample from the Scottish Mental Survey of 1947 (SMS1947, a cognitive ability test that included almost all Scottish children born in 1936) and linked it to the 1939 National Register, the National Health Service Central Register (NHSCR) and the SLS. The outcome of the project is a powerful life-course dataset containing information from childhood to old age.

The data linkage process has two distinct phases: (a) the automated phase where probabilistic matching was used to link data between the SMS1947 sample, the 1939 Register sample and the SLS, and (b) the manual phase where the clerical staff carried out manual linkages of the unmatched SMS1947 records to either the 1939 Register or the SLS (linking to the SLS was achieved by tracing the sample at the NHSCR database). Because the data collected in the 1939 Register are not fully computerised, the manual process also involved transcribing the 1939 registration data for the cohort members and their household members.

The quality of linkage obtained was extremely good. The success rate in linking the SMS1947 sample to the 1939 Register was 95.2%, after excluding those who were known to have migrated into Scotland after the 1939 registration date. The success rate in tracing the SMS1947 sample at NHSCR was 96.7%. The success rate in linking the traced SMS1947 sample to the SLS was 86.8%, after excluding those who were known to have died or emigrated. Further analysis of these tracing and linkage rates by gender revealed that they remained consistent between male and female groups. These high linkage success rates demonstrate the reliability of the linkage system being used.

Acknowledgements

This project is funded by the Medical Research Council as part of a multi-disciplinary programme, Lifelong health and wellbeing of the 6-Day sample of the Scottish Mental Survey 1947 (grant G1001401). The authors would like to thank Neil Bowie of NRS for undertaking automated record linkages, and the clerical support team from NHSCR for carrying out manual match and data input.

The SLS is supported by the Economic and Social Research Council Census programme, the Chief Scientist Office and the Scottish Government. The authors are responsible for the interpretation of the data. Census output is Crown copyright and is reproduced with the permission of the Controller of HMSO and the Queen's Printer for Scotland.

Contents

1 Background

The Scottish Longitudinal Study (SLS) is a contemporary longitudinal study which links data from the 1991, 2001 and 2011 censuses, vital events (births, deaths, marriages), National Health Service Central Register (NHSCR) data (migration in or out of Scotland), NHS data (cancer registry, hospital admissions, maternity data, etc.), pollution and weather data and educational data (school census and exam results). It is a 5.3% sample, based on 20 birth dates, providing linked information for approximately 274,000 individuals in Scotland.[1] Since its launch in 2007 the SLS has been used for a wide range of demographic, health and social-economic research (for a list of SLS projects see http://www.lscs.ac.uk/projects).

While we know much about the SLS members' lives from 1991 onwards, we know little about them prior to this date. The SLS Development and Support Unit (SLS-DSU) undertook a pilot project between 2008 and 2011, which investigated the feasibility of extending the SLS back through time via linkage to the 1939 Register and historical vital events registrations for both the SLS members and their parents.[2] The success of this pilot study has led to several database development projects aiming at adding value to the existing resources. One of them is linking the SLS with the Scottish Mental Survey of 1947 (SMS1947). The project took the SLS birth date sample from the SMS1947 (all members were born in 1936) and linked it to the 1939 Register, the NHSCR and the SLS. The outcome is a new SLS 1936 Birth Cohort Study (SLSBC1936) created from routine and administrative data, which contains information from childhood to old age.

The structure of the paper is as follows. Section 2 briefly describes the SMS1947, the 1939 Register and other NHSCR historical records. Section 3 sets out the objectives of the project, followed by the description of the process and methods used to link data between the SMS1947, the 1939 Register, the NHSCR and the SLS. Section 4 examines the quality of linkages between these datasets and compares the linkage rates with those achieved in the contemporary longitudinal studies. Section 5 summaries the datasets currently held in the SLSBC1936. The conclusions are given in Section 6.

2 Description of the Data

2.1 The 1947 Scottish Mental Survey

The SMS1947 was conducted by the Scottish Council for Research in Education. The survey designed to test the intelligence of a complete age-group of Scottish eleven-year-olds. It was intended to repeat, and for its results to be compared with, the Scottish Mental Survey of 1932. On 4th June 1947 almost all children who were born in 1936 and attending Scottish schools completed the same test that had been used in the Scottish Mental Survey of 1932. Those missing were mainly the absentees on the day of test. The test was a version of Moray House Test No. 12. It had 71 items, consisting of a variety of types of mental task: following directions (14 items), same–opposites (11), word classification (10), analogies (8), practical items (6), reasoning (5), proverbs (4), arithmetic (4), spatial items (4), mixed sentences (3), cipher decoding (2), and other items (4). A score of 76 was the maximum possible in the Moray House Test. Scotland is the only

country in the world to have tested mental ability in an entire population of people born in the same year.[3]

Two random sample of the SMS1947 were chosen to give further details: the 36-day sample and the 6-day sample – the latter being part of the former. A huge amount of data was collected on the 6-day sample via annual interviews from 1947 to 1963. A number of birth cohorts were created from the follow-up studies of the SMS1947, for example the Lothian Birth Cohort of 1936 which began in 2004 with the aim of discovering a wider range of causes of people's differences in cognitive ageing.[4] Recently the surviving members of the 6-day sample were recruited and comprehensive new data were collected as part of a multi-disciplinary project 'Lifelong health and wellbeing of the 6-Day sample of the Scottish Mental Survey 1947' led by the University of Edinburgh Centre for Cognitive Aging and Cognitive Epidemiology (CCACE).[5]

The SMS1947 database was held by the CCACE. It includes the following variables: pupil's surname, forename, other name, date of birth, sex, Moray House Test score, number of children in family, position in family and the name and locality of the school.

2.1 The 1939 National Register

The National Register was created shortly after the beginning of World War II under the National Registration Act 1939 and was an emergency measure which took a snap-shot of the entire UK population on 29th September 1939. It was used to provide information for issuing national identity cards, issuing food and clothing ration books, identifying children eligible for evacuation from areas vulnerable to bombing, and identifying adults eligible for call up into the Armed Forces. Subject to the specific exception of certain classes under the Act, registration was compulsory in respect of all members of the population present on National Registration Day. The excepted classes consisted persons serving in and not on leave from the Armed Forces, and people on ships in or nearing port.

In Great Britain the enumeration was planned, generally, on census lines. Following census procedure, the whole country was divided into more than 65,000 enumeration districts, each district being separately plotted with specified boundaries and specified street and house contents. The enumerators across the country delivered schedules ahead of National Registration Day. The individual returns were made on household schedules, and the responsibility for preparing the schedule was placed upon the head of household or, in the case of an establishment such as a hotel or institution, upon the manager or resident officer in charge, each of whom was required to supply the relevant record of every person who spent the night at the premises or arrived there the following day without having been enumerated elsewhere. The collection of the completed schedule was begun following National Registration Day. On receiving a schedule, the enumerator was required to check it and, if satisfied, issued a completed Identity Card for each person on the schedule. The enumerator was also required to group the returns in his district according to the address of the household and copy the schedules into a special transcript book. Checking the transcription books was carried out by local registration officers before their dispatch to the central offices.[6]

When the National Health Service Central Register (NHSCR) was formed in 1948, the National Register was adapted for both Registers, and the National Registration number (a combination of the enumeration district (ED) code, household number and person number) became the NHS number. In February 1952 National Registration was discontinued and the Identity Card was abolished, but the records have continued to be used by NHSCR. The Registers were turned over to their present role in the administration of the healthcare system. The historical NHSCR registers contain eight different volume types (see Table 1). In total there are 2132 volumes of manuscript records. These resources were used and maintained by NHSCR before the computer database went live in 1986. In 2008 National Records of Scotland (NRS) completed the project of creating digital images of all historical NHSCR registers. An electronic index was also produced allowing searches to be conducted and a reference established which linked to the relevant images.

Information originally collected in the 1939 Register includes: address, ED code, household number, person number, surname, forename, sex, date of birth, marital status and occupation. This information was regarded as necessary for the purpose of the National Register. Other information updated or amended by NHSCR includes postings which recorded the moves of an individual subsequent to 1939 up to the late 1950s, changes of name or marital status, and date of death if the person died. Because the data are grouped into households it is possible to identify household information such as relationships, number of persons, adults and children in the household.

Table 1: Types of historical NHSCR Registers

Ref	*Type*	*Description*	*No of volumes*
1	National Register	Transcript entries of all persons enumerated in Scotland on 29/9/1939.	1224
2	New Entrants to National Register	Persons entering Scotland from outside the UK from 1939-1952.	107
3	Demobalisation Register	People released from the forces between the end of the war and 1952	136
4	Lost Identity Card Register	Recorded new numbers to those who lost their identity/ration cards during 1940-1948	47
5	Seamen & Vagrants Register	Persons missed in the original enumeration in 1939 but were registered on entering a Scottish port or picked up as a vagrant	1
6	Scottish Central Register	Started at the end of National Registration for immigrants or persons whose original numbers could not be traced. This register was in existence from 1952 until 1972.	89
7	Allocated Number Register	This register replaced the Scottish Central Register in 1972.	26
8	Birth Register	Records births of all people born in Scotland after 29/9/1939	502

The computerised information in the 1939 Register index includes: surname, forename, sex, date of birth, ED code, household number, person number, volume number and image number. A separate index entry was created for each amended entry due to a change of name, one to capture the original information

and the other to capture the revised information. Separate index entries were created for all combinations of a hyphenated surname. Therefore, although the transcription books contain a single line entry for each individual, one person may have multiple entries in the index.

3 Objectives and Methods

3.1 Objectives

The aim of this project is to create a new 1936 Birth Cohort Study (i.e., the SLSBC1936) through longitudinal linkages between the SLS, the SMS1947, the 1939 Register and the NHSCR database. The cohort was drawn from the SMS1947 dataset using the SLS birth dates. The SMS1947 dataset contains 75,252 records, of which 4,165 records have one of the SLS birth dates. After removing the duplicate records the SLSBC1936 contains 4,162 unique members. The objectives of this project are:

- to link information from the 1939 Register to the cohort members who were enumerated in Scotland in 1939 (i.e., to link the SMS1947 with the 1939 Register);
- to link information from the existing SLS dataset to the cohort members who were captured in the SLS since 1991 Census day (i.e., to link the SMS1947 with the SLS); and
- to trace and flag the cohort members in the NHSCR database (i.e., to link the SMS1947 with the NHSCR).

Since the SLS members are flagged in the NHSCR database, the cohort members successfully linked to the SLS would have been traced and flagged in the NHSCR (a summary of tracing the SLS members, flagging and linking data to them is given in LSCS Working Paper 1). Thus the third linkage only applies to those members who have not been linked to the SLS. Once the cohort members are traced in the NHSCR, their health data will be linked through the routine SLS – ISD link.

3.2 Methods of linkage

The creation of the SLSBC1936 was a complex process. It involved linking data across multiple sources and collected at different time points. The linkage process was divided into two distinct phases: the automated phase and the manual phase. In the automated phase we focused on the linkages of three datasets extracted from the SMS1947, the 1939 Register index and the SLS. The unmatched SMS1947 records were passed to the manual phase where a cohort member was matched against the full 1939 Register or the NHSCR database. Because the data collected in the 1939 Register are not fully computerised, the manual processes also involved identifying the relevant records and entering in the information missing from the partial transcriptions of the index.

Because the data held in the SLS are anonymous, data linkage was carried out outside the SLS setting. For the purpose of this study the CCACE provided the SMS1947 data. NRS provided the 1939 index data and undertook automated linkage. Manual linkage and transcription of the 1939 registration data were undertaken by the clerical support team from NHSCR. The SLS-DSU designed the

linkage process, provided the interactive systems for data entry and manual matching, reviewed the linkage results and created the linked dataset for researchers. The steps involved in the linkage process are described below.

Step 1: Preparing data extracts for auto-matching

NRS prepared the following three data files for automated process. Each file includes the variables available for data matching.

- File A contains 4162 cohort members (i.e, the 20-day sample of SMS1947). Each individual was allocated a unique record identifier (SLS 1947 entry number). The information in this file includes: surname, forename, sex, date of birth, locale of school and county.
- File B was extracted from the index of 1939 Register. It contains all members born in 1936 with a SLS birthday. This file contains 6275 records which refer to 4467 persons according to the 1939 registration number. Information in this file includes: surname, forename, sex, date of birth and the 1939 registration number (i.e., the original NHS number).
- File C was extracted from the NHSCR database. It contains the SLS members traced at NHSCR who were entered the SLS project up to and including the 2001 Census date and born in 1936 (N = 3313). The information in this file includes: SLS number, NHS number, surname, forename, previous name, sex and date of birth.

The format of the NHS number in the NHSCR dataset (File C) is irregular. NHS numbers originally issued consist of a 4-letter ED code followed by household number (schedule number) and person number (sub-number). New entrants to the National Register were allocated a number consisting of a 3-letter code (identifying each administration area) and a serial number of up to 6 digits. Various other types of NHS number issued later. In addition, SLS members who had moved from England and Wales into Scotland would have an England/Wales NHS number format. In order to compare the NHS number between Files B and C the format of NHS numbers in File C was standardised by splitting the NHS number into separate components and removing any separators. Then the sources of NHS numbers were analysed and each record was classified into one of NHS number types.

Step 2: Auto-matching

Three data matching exercises were conducted to link between the three data files created in Step 1. Table 2 summaries the methods and matching variables used in each linkage exercise. Probabilistic record linkage method was employed in all exercises. The probabilistic record linkage software used in this project is Link Plus, which is a free, stand-alone application for Microsoft Windows.[7] The program generates a probabilistic record linkage score that indicates, for any pair of record, how likely it is that they both refer to the same person. Manual review was used to set the lower and upper cut-off thresholds. Any linked pair with a score above the upper cut off value is regarded as a match. Any linked pair with a score under the lower cut off value is regarded as a non-match. Pairs with score between the thresholds are possible matches, and they were clerically inspected by the data linker.

Table 2: Automated processes: method and matching variables

Type of linkage	*Method*	*Matching variables*
1. File A-File B (SMS1947-1939 Index)	Probabilistic match	surname, forename, date of birth and sex
2. File A-File C (SMS1947-SLS)	Probabilistic match	Surname/previous surname, forename, date of birth and sex
3. File B-File C (1939 Index-SLS)	Direct match	NHS number and date of birth, or NHS number and forename
	Probabilistic match 1	surname, forename, NHS number, date of birth and sex
	Probabilistic match 2	previous surname, forename, NHS number, date of birth and sex

SMS1947 (File A) – 1939 Index (File B) Link
This linking was achieved on the basis of surname, forename, date of birth and sex. We found out the combination of these variables is not unique in the SMS1947 sample. File A contains 22 pairs of records with the same surname, forename, date of birth and sex. These records were visually examined to see if a match could be made using the geographic information in the two datasets, for example comparing the 1939 registration district and the 1947 school locality. They were also flagged for more rigorous checks at the manual match stage.

SMS1947 (File A) – SLS (File C) Link
This linking was also achieved on the basis of surname, forename, date of birth and sex. Unlike date of birth and sex which are fixed at birth and rarely changes, person's names may change during a person's lifetime, for example, when women adopt their husbands surname on marriage. Fortunately the NHSCR database also includes previous surname. Therefore for this exercise surname in the SLS was replaced with previous surname where it was possible. Those records with the same combination of surname, forename, date of birth and sex in the SMS1947 sample were flagged for manual processing.

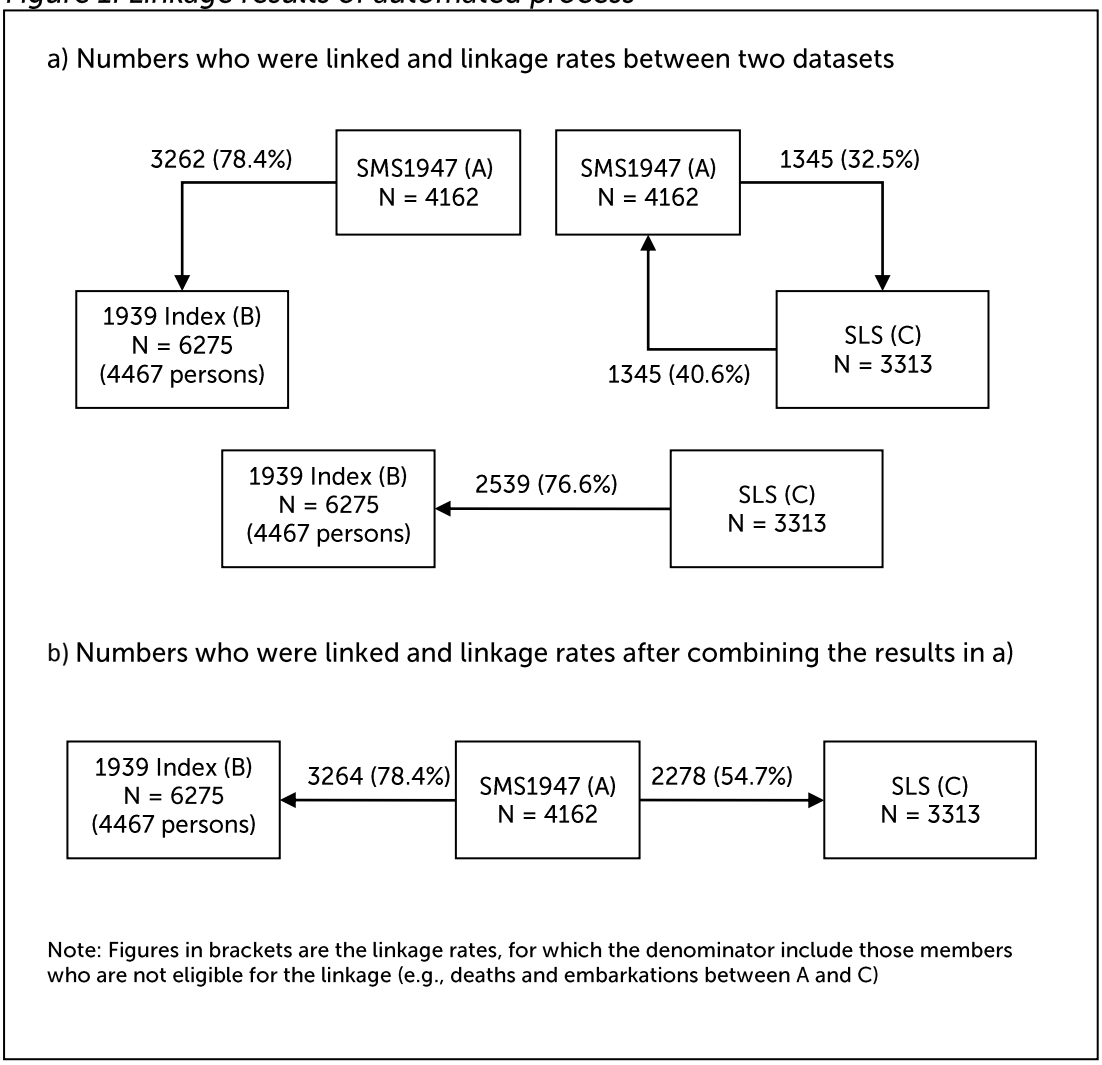
SLS (File C) – 1939 Index (File B) Link
A combination of direct (exact) match and probabilistic match was employed in this exercise. Because both files contain the NHS number direct match was carried out before the probabilistic match. Direct matching was made using two sets of variables (1) NHS number and date of birth, and (2) NHS number and forename. If the conditions in any set are met, the record pair is considered a direct match. In order to maximise possible links the probabilistic matching was run twice, using current surname then previous surname together with forename, date of birth, sex and NHS number. Note that this part of work was completed during the feasibility study of linking the SLS backing through time where the full SLS dataset, including all SLS members born before the 1939 registration day, was matched against the SLS day sample of 1939 index. Files B and C are the subsets of these linked datasets; hence the list of matched records is already available.

The cohort members who came into Scotland from abroad after the 1939 registration day would not be eligible for the backward linkage to the 1939 Register. In order to identify these members an additional probabilistic match was carried out to link those members in the SMS1947 who have not been found in the 1939 index to the index of New Entrants to National Register.

Step 3: Processing the auto-matching results

The link files created in Step 2 were passed to the SLS-DSU for processing. This step involved checking the multiple matches and inconsistencies, and combining the three separate link files into a single link table. Multiple matches were generated when linking the 1939 index with the SMS1947 or the SLS because one person could have multiple entries in the 1939 index. Where several records in the 1939 index were linked to one SMS1947 (or SLS) record, their 1939 registration member (NHS number) must be identical otherwise the links were rejected. When we analysed the linkage results and combined the three link files into a single link table the 1939 registration number was used as the *code* by which individuals are identified in the 1939 Register (File C).

Figure 1: Linkage results of automated process



The results of auto-matching are illustrated in Figure 1. Part (a) shows the numbers who were linked and the linkage rates of each linkage exercise and part (b) shows the numbers who were linked and the linkage rates when the results of separate links were combined (e.g., a record in A is linked to a record in C either by direct linking between A and C or by linking A with B and B with C). At the end of this stage a link table was created, which shows the linkage status of each cohort member. Table 3 illustrate the structure of the link table. The cohort members were classified into four groups. Group A contains those SMS1947 records which have been linked to a 1939 index record and a SLS record. Group B contains those SMS1974 records which have been linked to a 1939 index record but have not been linked to the SLS. Group C contains those SMS1974 records which have been linked to a SLS record but have not been linked to the 1939 index. Group D contains those 1947SMS records which have neither been found in the 1939 index nor in the SLS.

Table 3: Structure of the link table

Individual	*Group*	*SMS1947 code (File A)*	*1939 Index code (File B)*	*SLS code (File C)*
1	A	47001921	SAAA/495/2	10100231
2	B	47004586	SAAB/35/5	
3	C	47008693		10000002
4	D	47001056		

Step 4: Developing a system for manual process

A comprehensive software tools called *SLSLINK1947* was developed in house. The system was designed in Microsoft Access and consists of two parts: an Access database for storing the data and a front-end system for linking and inputting data. Because the tasks of manual exercise change according to the outcome of auto-matching, the system includes four screens especially designed for processing records in Groups A, B, C and D. Each screen includes a main form and a sub-form. The main form shows the member's information collected in the SMS1947 and the linked 1939 index data or the SLS data where possible. It allows the results of automated linkage to be updated. The sub-form provides a utility for capturing the 1939 data for the cohort member and other household members.

Table 4: Number of records in each group and tasks of manual exercise

Group	No of SMS1947 records (File A)	Linked to a 1939 Index record (File B)	Linked to a SLS record (File C)	NHSCR main tasks
A	2086	Yes	Yes	• Transcribe 1939 data
B	1178	Yes	No	• Transcribe 1939 data • Trace the record at the NHSCR
C	187	No	Yes	• Match the record against 1939NR • Transcribe 1939 data if a match is found
D	711	No	No	• Match the record against 1939NR and input 1939 data if a match is found • Trace the record at NHSCR

Step 5: Manual matching

NHSCR carried out manual linkage of the unmatched SMS1947 records to either the 1939 Register or the SLS and transcribed the 1939 registration data for those who were linked to the 1939 Register. In this phase the unmatched SMS1947 records were matched against the master NHSCR database or the full 1939 Register. For each record data linkage and transcription of 1939 data were carried out simultaneously starting from group A to group D. Table 4 shows the number of records in each group and the main tasks of manual exercise.

The specific tasks include:

- Manually review the SMS1947 records which have not been found in the 1939 Register or in the SLS in the auto-match phase.
- Flag the cohort members on the NHSCR database if they are not linked to an existing SLS member.
- Trace the cohort members at England and Wales NHS system if they are not found in the NHSCR database.
- Check and correct any linkage errors in the auto-matching phase.
- Transcribe relevant data from the 1939 Register for those who were linked to the 1939 Register (see below).
- Identify deaths occurring to the cohort members before 1991 from the NHSCR database.
- Identify migrants out of Scotland before 1991.

Searching the 1939 Register images and transcribing the data benefited from the auto-matching. Firstly, once a cohort member is linked to a 1939 Register index, the 1939 data (names, sex, date of birth, household number and person number) of the cohort member and their household members can be extracted from the 1939 index. As a result we do not need to re-enter this information. Secondly, the reference to the image file (volume number and image number) would be

available from the linked 1939 Register index, which provides faster access to the 1939 Register image compared with the use of search engine.

The following information was captured from the 1939 Register:

- Address
- Record type (private household or communal establishment)
- Household size
- Type of communal establishment
- Marital status
- Occupation
- Relationship to SLS member

For the cohort members enumerated in a household all household members' data were transcribed. The household relationships were not recorded in the 1939 Register. NHSCR coded *the relationship to SLS member* based on the household member's name, sex, date of birth and marital status. In some cases the cohort member's birth certificate was checked to confirm the relationship.

Step 6: Post-process

This step included checking the data quality, coding 1939 occupation titles and addresses, linking the data to the main SLS dataset and creating a SLSBC1936 research database. On completion of step 5, an anonymised linked dataset with the names removed was passed to the SLS-DSU. All records linked to the NHSCR database were checked against the latest NHSCR extract supplied to the SLS to ensure that the study reference number and the SLS numbers of new entries were correctly entered in the NHSCR database. Unmatched records were sent to NHSCR for reconciliation.

Transcribed data were checked and any missing data were updated. Migration data (year of leaving Scotland) and death data (date of death) before 1991 Census were manually entered by NHSCR. They were also checked against the latest NHSCR migration data supplied to the SLS; any data missed at manual match stage were manually amended.

The 1939 occupation titles were manually coded to the Historical International Standard Classification of Occupations (HISCO). For geocoding, we used the 'Historical Address Geocoding – GIS' (HAG-GIS) software package that was developed for geocoding historical data. The matching algorithm links the historical addresses to the contemporary addresses by exact and probability matching methods. Manual matching is also used to complement the process for unmatched addresses.[9]

Finally, a database containing the SLSBC1936 core information, the SMS1947 data and the 1939 data was created, which is linked to the main SLS dataset via the SLS number.

4 Quality of the linkage

The quality of the SMS1947 – 1939 Register linkage can be measured by looking at the backward linkage rate, i.e., how many people in the SMS1947 should have been in the 1939 Register compared with those who were found in the 1939 Register. The SMS1947 members who migrated into Scotland after 1939 registration date were excluded. The quality of the SMS1947 – SLS linkage can be measured by looking at the tracing rate of the SMS1947 sample at NHSCR and then the forward linkage rate between the traced SMS1947 sample and the SLS, given that only those SMS1947 members who were traced in the NHSCR could be linked to the SLS. The forward linkage rate between the SMS1947 and the SLS is based on those numbers who were known still alive at 1991 Census, and who were not recorded as having emigrated during the period between 1947 and 1991. Figure 2 shows the cohort numbers who were linked or not to the 1939 Register, the NHSCR and the SLS and the overall linkage success rates. In the following sections these linkage rates are examined by sex as recorded in the SMS1947.

4.1 Backward linkage rate between SMS1947 and 1939 Register

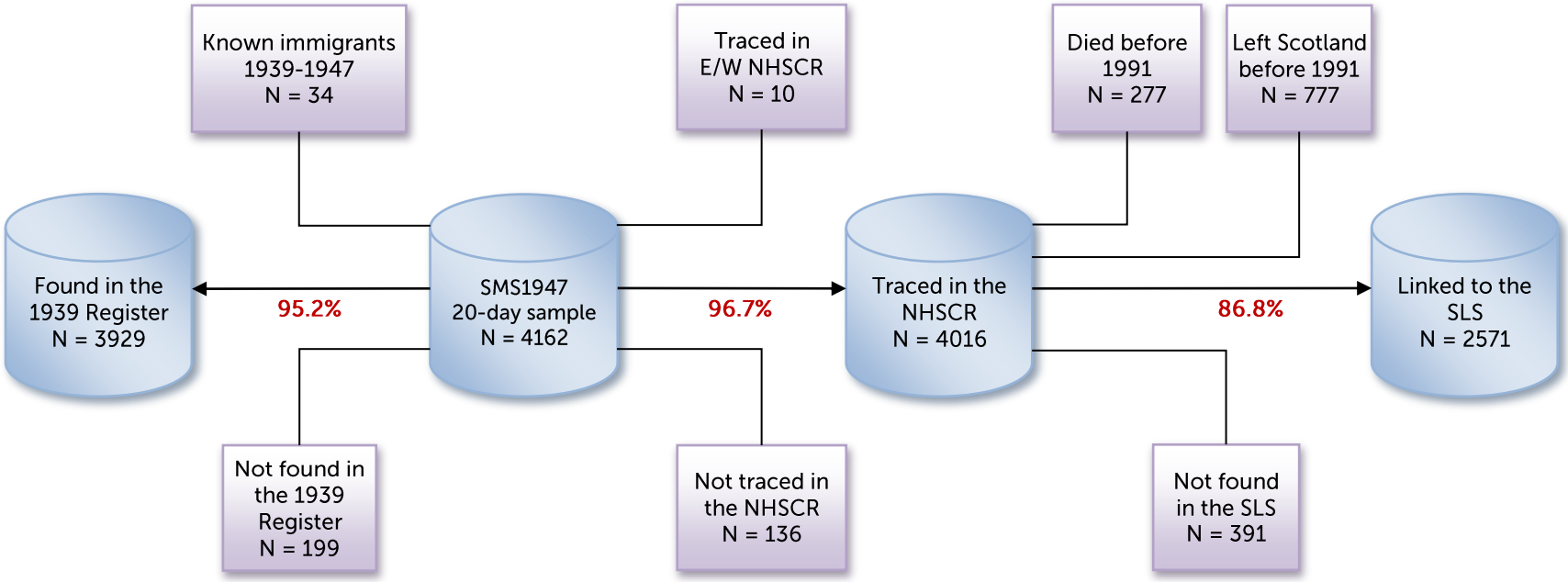
Table 5 shows the cohort's backward linkage rates between the SMS1947 and the 1939 Register by sex. The cohort was made up of 2113 males and 2049 females. 14 males and 20 females were known immigrants because they were either found in the New Entrants to National Register or enumerated in England and Wales in 1939 (identified by their NHS number). Thus 2099 males and 2029 females were eligible to link back to the 1939 Register, of which 2000 (95.3%) males and 1929 (95.1%) females were successfully linked. The overall backward linkage rate was 95.2%.

Table 5: Backward linkage rates between SMS1947 and 1939 Register by sex

	Male	*Female*	*Total*
20-day sample of SMS1947	2113	2049	4162
Known immigrants between 1939 and 1947*	14	20	34
Should have been in the 1939 Register	2099	2029	4128
Found in the 1939 Register	2000	1929	3929
Linkage success rate	*95.3%*	*95.1%*	*95.2%*

** These are people found in the new entrants to National Register or enumerated in England and Wales in 1939.*

Figure 2: Longitudinal linkages of the 20-day sample of SMS1947 to the 1939 Register, the NHSCR and the SLS



4.2 Tracing rate at NHSCR

Table 6 shows the success rates in tracing the cohort members at NHSCR by sex. 5 males and 5 females in the SMS1947 sample were not traced at NHSCR but were found in England and Wales NHS system. These people never registered with a Scottish doctor; hence they were not eligible to be traced in the NHSCR. Of 2108 males and 2044 females who were expected to be traced in the NHSCR, 2046 (97.1%) males and 1970 (96.4%) females were actually traced. Overall, the tracing rate was 96.7%.

Table 6: Tracing rates of the SMS1947 members at NHSCR by sex

	Male	*Female*	*Total*
20-day sample of SMS1947	2113	2049	4162
Residents in England and Wales*	5	5	10
Eligible to be in the NHSCR	2108	2044	4152
Traced at NHSCR	2046	1970	4016
Tracing rate	*97.1%*	*96.4%*	*96.7%*

** These are people only traced at England and Wales NHS system, but never found in the Scottish NHSCR.*

4.3 Forward linkage rate between SMS1947 and SLS

Table 7 shows the forward linkage rates between the traced SMS1947 sample and the SLS by sex. Of the cohort members traced at NHSCR 173 males and 104 females died before 1991 Census, and 425 males and 351 females left Scotland before 1991 Census and did not returned prior to 2001 Census. This left 1448 male and 1514 female traced at NHSCR who were not eligible to be linked to the SLS, of which 1262 (87.2%) males and 1309 (86.5%) females were successfully linked to the SLS. The overall linkage rate was 86.8%.

Table 7: Forward linkage rates between SMS1947 and SLS – sample members traced at NHSCR by sex

	Male	*Female*	*Total*
20-day sample of SMS1947 traced at NHSCR	2046	1970	4016
Died before 1991 Census	173	104	277
Embarked before 1991 Census	425	351	777
Eligible to be in the SLS	1448	1514	2962
Linked to the SLS	1262	1309	2571
Linkage success rate	*87.2%*	*86.5%*	*86.8%*

4.4 Comparison of the linkage quality with the LS Census link

The ONS Longitudinal Study (LS) links census, vital events data and NHS registration information for 1% of the population of England and Wales. It was started in 1974 with the initial sample drawn from the 1971 Census. Subsequent sample have been drawn and linked from the 1981, 1991, 2001 and 2011 Censuses using the LS dates of birth. Linkage of data over time is achieved by

tracing against NHS patient records. Because the data linkage quality is extremely high the LS has been used as the benchmark for assessing the linkage quality of other census-based longitudinal studies. Table 8 shows the linkage success rates between two successive LS-Census samples traced at NHSCR.[8]

Table 8: Linkage rates between two LS-Census samples traced at NHSCR

	Forward linkage rate (%)	*Backward linkage rate (%)*
1971 – 1981 Link	91.3	92.6
1981 – 1991 Link	90.1	91.4
1991 – 2001 Link	88.0	90.7
2001 – 2011 Link	87.7	90.2

Notes:
1. Source: ONS Longitudinal Study for England and Wales.
2. In 2011, census data were traced against the new Medical Research Information Service Integrated Database Administrative System (MIDAS)

The backward linkage rate of 95.2% between the SMS1947 and the 1939 Register was extremely high when compared to the linkage rates reported in the LS Census link, even taking into account the fact that the gap between the two datasets is less than 10 years. It should be noted that only traced LS members, for whom a record has been found on the NHS system, were included in the assessment of linkage rates. If we used the same definition the backward linkage rate between the SMS1947 and the 1939 Register for the cohort members traced at NHSCR would be a remarkable 98.4%.

To assess the quality of the SMS1947 – SLS link we can compare the tracing rate of the SMS1947 sample at NHSCR and the forward linage rate between the SMS1947 and the SLS with those achieved in tracing and linking the LS 1971 members to the LS 2011 Census sample for the same age cohort. The ONS-LS development team provided the data for this comparison. There were 7,792 persons in the LS1971 11-year-old cohort, 7,616 (97.7%) of them were traced at NHSCR. Note that tracing the LS 1971 census sample at NHSCR was carried out between 1974 and 1976 and it was purely a clerical process. Whilst tracing the SMS1947 sample at NHSCR was conducted over six decades later, the tracing rate of 96.7% obtained in this project was therefore extremely good.

The forward linkage rates between 1971 and 2011 censuses for the LS1971 11-year-old cohort traced at NHSCR was 87.6% (Table 9). This was very impressive given that the gap between the two datasets was 40 years. The success rate of 86.8% obtained in linking the SMS1947 to the SLS was slightly lower, but the gap between the SMS1947 and the SLS was 44 years. Therefore the quality of the SMS1947 – SLS link is comparable with the LS census link.

It should be noted that the out-migration rate of the SLSBC1936 was much higher than that of the LS1971 11-year-old cohort. 19.4% of the traced SLSBC1936 members left Scotland during the period between 1947 and 1991, while only 3.8% of the traced LS1971 11-year-old cohort embarked before 2011 Census. Identifying emigrants from the UK relies on study members informing their GP when they are leaving the country, which many fail to do so. The success rate in linking data to the SLS was more likely to be affected by the missed migrations out of the country due to the high out-migration rate. For example,

the overall forward linkage rate of the SLS 1991 – 2001 Census link was 85.6% compared with 88.0% for the LS 1991 – 2001 Census link, although the linkage methods employed by the two studies were similar. To achieve a success rate of 86.8% in linking the SMS1947 to the SLS was truly remarkable. The fact that the linkage rate between the SMS1947 and the SLS was even higher than the overall forward linkage rate between 1991 and 2001 SLS Census samples further confirmed that that linking the SMS1949 sample to the SLS was extremely successful.

Table 9: Forward linkage rates between 1971 and 2011 Censuses (LS1971 11-year-old cohort)

	Total
Present in 1971 Census aged 11 traced at NHSCR	7,616
Died before 2011 Census	307
Embarked prior to 2011 Census	290
Eligible to be in 2011 Census	7,019
Found in 2011 Census	6,149
Forward linkage rate	*87.6%*

Source: ONS Longitudinal Study for England and Wales

The unlinked cases were probably due to missed events (emigration, death), date of birth discrepancies between the datasets involved, changes of name and people who were not enumerated at Census. The problem of capturing accurate historical migration data was probably a major cause of any failure in linking the SMS1947 to the SLS.

5 Data included in the SLSBC1936

The datasets currently included in the SLSBC1936 are summarised in Table 10. Basically, they include the 1939 Register, the SMS1947, Censuses (1991-2011), Vital Events data (birth, deaths and marriages), NHSCR data and health data. The 1939 Register provides the family and household information of the cohort at age 3 years, including the age and occupation of the cohort's parents, the relationship of other household members to the cohort member, and the location of household or communal establishment. The SMS1947 provides information on the cognitive ability of the cohort at age 11 years. The 1991, 2001 and 2011 Census data include a wide range of demographic, cultural, health, housing, employment and social variables. It provides information on the cohort's final achieved education qualification, occupation, mid-to-late life household composition and social and geographical position at ages 55, 65 and 75 years. Self-reported health status was included in 2001 and 2011 Censuses and a new question on long-term health condition was added to 2011 Census. These data are also available for those living in the same household as a cohort member.

Vital events data collected for the SLSBC1936 include births to the cohort members, deaths of the cohort members, widow(er)hoods of cohort members (where the cohort member is the surviving spouse) and marriages of the cohort

members. They are from the existing SLS database (vital events data for the SLS members are linked to and held on the SLS database). These data have been added for the period 1991-2011. Vital events data after 2011 will be added through the SLS annual update exercise. It is also planned to include the cohort's death events between 1974 (when the information was first collected electronically) and 1991. The NHSCR data provides information on the migration history of the cohort from 1948 onwards, including moving across health boards, emigration out of Scotland and re-entries into Scotland after previous emigrations. It also records all the deaths occurring to the cohort members since 1948, although only date of death and registration details has been captured together with a country of death indicator.

Health data are provided by the Information and Services Division (ISD) of the NHS Scotland. Health data are not held on the SLS database, but are linked on a project-by-project basis. ISD holds a SLS/ISD lookup table containing encrypted SLS identifiers. This allows the linkage of specific data to SLS members if permission is granted to use that data in a particular study. So far health data are available for the cohort members who were linked to the SLS because the existing SLS/ISD lookup table was created before this project started. We are currently updating the SLS/ISD lookup table to add the cohort members who were not linked to the SLS but traced at NHSCR, including those who died or embarked before 1991 Census.

Detailed Information on the variables relating the SLSBC1936 can be obtained by contacting the SLS-DSU (email sls@lscs.ac.uk). Information on the variables in the SLS is available in the SLS data dictionary that can be found online at http://www.sls.lacs.ac.uk/variables/.

Table 10: Data currently held in the SLSBC1936

1939 National Register	Age, sex, marital status
	Family, household or communal establishment type
	Occupations
	Geo-coded address
Scottish Mental Survey 1947	Cognitive ability: Moray House Test
	Family
	Locality of school
1991 Census	Age, sex, marital status
	Family, household or communal establishment type
	Housing, including tenure, rooms and amenities
	Country of birth
	Ethnicity
	Educational qualifications
	Economic activity
	Occupation and social class
	Migration
	Limiting long-term illness
	Geo-coded address
2001 Census	Similar data to 1991, but additional information collected in 2001 includes:
	self-rated health
	religion
	caregiving
2011 Census	Similar data to 2001, but additional information collected in 2011 includes:
	national identity
	long term health conditions
	language
Vital Events (1991 – 2011)	Births to sample mothers and fathers
	Death of sample members
	Widow(er)hoods of sample members
	Marriages of sample members
NHSCR data (1948 onwards)	Migration into and out of Scotland
	Changes of health board
	Date of death
	Registration details
	Country of death (Scotland/England and Wales/Abroad)
Health data	Hospital admissions (SMR01)
	Cancer Registrations (SMR06)
	Maternity Inpatient and Day Case (SMR02)
	Mental Health Inpatient and Day Case(SMR04)
	Outpatient Attendance dataset (SMR00)
	Prescribing Information System (PIS) (2009 onwards)

1939 Register and Census data are available for the cohort members and those living in the same household as a cohort member.

6 Conclusion

This paper presents a pragmatic approach of retrospectively constructing longitudinal birth cohort studies from already existing datasets. The SLSBC1936 was created through longitudinal linkages of the 20-day sample of the SMS1947 to the 1939 Register, the NHSCR and the SLS. Data linkages were extremely successful with 96.7% of the cohort members were traced at NHSCR, 95.2% of the cohort members were found in the 1939 Register after excluding those who were known to enter into Scotland after the 1939 registration date, and 86.8% of the traced cohort members were linked to the SLS after excluding those who were known to have died or who had emigrated before 1991. Further analysis of these tracing and linkage rates by gender revealed that they remained consistent between male and female groups. The problem of capturing accurate historical migration data at NHSCR was probably a major cause of failure in linking the SMS1947 to the SLS.

The SLSBC1936 combines information from the 1939 Register, the SMS1947 and the SLS. It is expected that with a relatively small amount of funding and dedicated time, additional linkages such as marriages of the cohort and deaths occurring to the cohort members before 1991 Census could be completed. The SLSBC1936 is a general-purpose resource, which is available for researchers via the SLS administration. Anyone interesting in accessing this dataset should visit our website (www.sls.lscs.ac.uk) or contact the SLS-DSU at sls@lscs.ac.uk.

7 References

1. Boyle, P, F. Peteke, Z. Feng, L.Hattersley, Z. Huang, J. Nolan and G. Raab, Cohort Profile: The Scottish Longitudinal Study (SLS) *International Journal of Epidemiology*. 2008.

2. Boyle, P. and Z. Huang. Testing the feasibility of extending the Scottish Longitudinal Study back through time: *ESRC End of Award Report*, RES-348-25-0014: ESRC.

3. Scottish Council for Research in Education. The trend of Scottish Intelligence: A comparison of the 1947 and 1932 surveys of the intelligence of eleven-year-old pupils. London: University of London Press. 1949.

4. Deary, I.J., A. J. Gow, A. Pattie and J.M. Starr, Cohort Profile: The Lothian Birth Cohorts of 1921 and 1936, *International Journal of Epidemiology*. 2011

5. Deary, I.J. and C.E. Brett. Predicting and retrodicting between childhood and old age in the 6-Day sample of the Scottish Mental Survey 1947. *Intelligence*. 2015.

6. National Register United Kingdom and Isle of Man: Statistics of Population on 29th September, 1939. London, His Majesty's Stationary Office.

7. Thoburn K, Gu D, Rawson, T, Rogers, J. Link Plus Version 2: An Essential Central Cancer Registry Linkage Tool. NAACCR 2008 Annual Conference. Denver, Colorado. 2008.

8. Lynch, K., S. Leib, J. Warren, N. Rogers and J Buxton. Longitudinal Study 2001 – 2011 Completeness of census linkage. Office for National Statistics. Series LS No. 11.

9. Daras K, Feng Z, Dibben C and Williamson L, 2016 Digitising and geocoding historical vital events in Scotland from 1855 to 1974, ESSHC Valencia.