The creation of an administrative data based 1936 Birth Cohort Study

Zengyi Huang[1], Chris Dibben[2], Graham Kirby[3], Ian Deary [2], Frank Popham[4], Roxanne Connelly[2]

[1] LSCS, University of Edinburgh, [2] University of Edinburgh, [3] University of St Andrews, [4] University of Glasgow

Overview

The objective of this project is to create a new 1936 Birth Cohort Study from routine and administrative data. It is structured around the existing Scottish Longitudinal Study (SLS). We took the SLS birth date sample from the Scottish Mental Survey of 1947 (SMS1947) and linked it to the 1939 Register, the National Health Service Central Register (NHSCR) and the SLS. The outcome is a powerful life-course dataset containing information from childhood to old age.

Description of the Data

The SLS - The SLS is a contemporary longitudinal study which links data from the 1991, 2001 and 2011 censuses, vital events (births, deaths, marriages), 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. Further details can be found at http://sls.lscs.ac.uk/

The SMS1947 – The SMS1947 was conducted by the Scottish Council for Research in Education on 4th June 1947. Almost all Scottish schoolchildren born in 1936 at age 11 sat the same intelligence test, with scores recorded from 70,805 children. The SMS1947 database includes pupil's surname, forename, date of birth, sex, Moray House Test (MHT) score, number of children in family, position in family and the name and location of the school.

The 1939 National Register – The census like 1939 register was established shortly after the beginning of WWII to provide information for issuing national identity cards. Every person was given their own unique civil registration number, based on where they were living on 29th September 1939 and this became the NHS number when the NHS was created in 1948. The information collected in the 1939 register includes person's address, surname, forename, sex, date of birth, marital status, occupation, and postings which recorded the moves of all these individuals subsequent to 1939 up to the late 1950s. The 1939 register is available as digital images with a supporting index. The computerised information in the index includes: surname, forename, sex, date of birth, NHS number, image number, and reference to source.

Method of Linkage

The creation of the 1936 Birth Cohort was a complex process and involved collaboration between the National Records of Scotland (NRS), the University of Edinburgh Centre for Cognitive Aging and Cognitive Epidemiology (CCACE) and the SLS team. The CCACE provided the SMS1947 data. NRS provided the 1939 register 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 linkage process was divided into six steps.

1. **Pre-processing** – Selected the Cohort members and created three linkage files for automated process: (a) 20-day sample of SMS1947 (i.e., the Cohort members), (b) 20-day sample of the 1939 register index, and (c) SLS members traced at NHSCR and born in 1936.
2. **Auto-matching** – Three data linkage exercises were carried out. Table 1 shows the methods and matching variables used in each linkage exercise.
3. **Processing the auto-matching results** – The three matched datasets were combined into a single link file. Multiple links and inconsistencies were identified and flagged for manual review.
4. **Developing a system for manual process** – An integrated software tool was developed in house for manual matching and data entry.
5. **Manual matching** – NHSCR carried out manual linkage of the unmatched SMS1947 records to either the 1939 register or the SLS. The SMS1947/1939 register linkage involved identifying the relevant 1939 records and entering in the information missing from the partial transcriptions of the index for the Cohort members and other household members. The SMS1947/SLS linkage was made by tracing and flagging the Cohort members in the NHSCR database.
6. **Post-processing** – This step included checking the data quality, coding 1939 occupation titles and addresses, creating derived variables and linking the data to main SLS dataset.

Table 1. Automated process: methods and matching variables

	Method	Matching variables
Linking File (a) and File (b) (SMS194-1939 Index)	Probabilistic match	Surname, forename, date of birth and sex
Linking File (a) and File (c) (SMS1947-SLS)	Probabilistic match	Previous surname/surname, forename, date of birth and sex
Linking File (b) and File (c) (1939 Index-SLS)	Direct match	NHS number and date of birth or NHS number and forename
	Probabilistic match	Previous surname/surname, forename, NHS number, dob, sex

Quality of the Linkage

Figure 1 Longitudinal linking of the SMS1947 to the 1939 Register, the NHSCR and the SLS

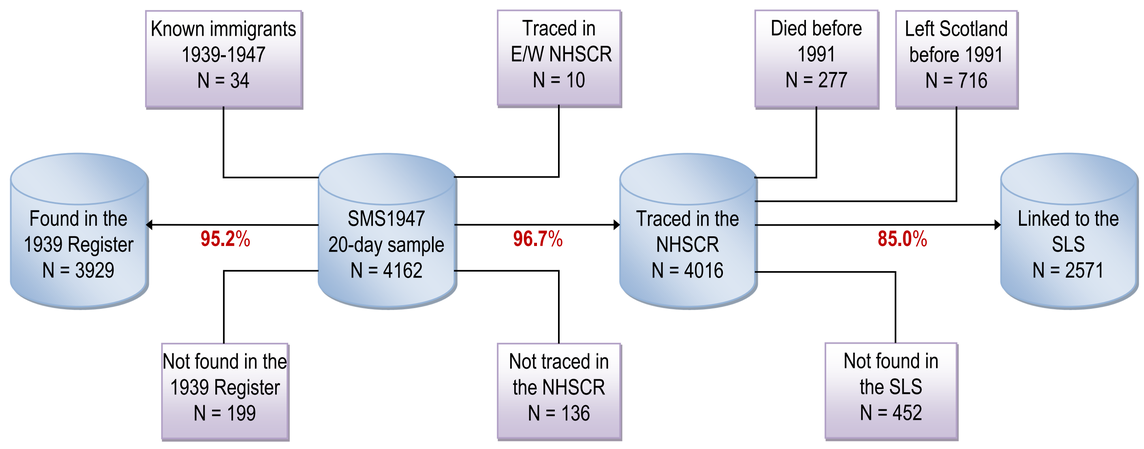


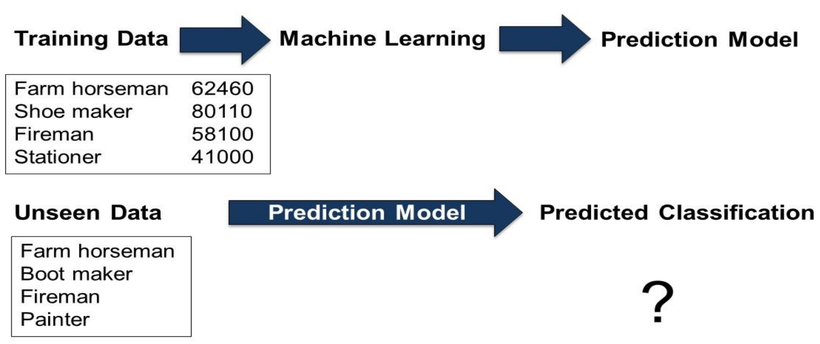
Table 2. Forward and backward linkage rates by sex

	Male	Female	All
Backward linkage rate between 1939 Register and SMS1947	95.3%	95.1%	95.2%
Forward linkage rate between SMS1947 and NHSCR	97.1%	96.4%	96.7%
Forward linkage rate between SMS1947* and SLS	85.0%	85.1%	85.0%

* Traced at NHSCR. Source: Scottish Longitudinal Study.

Coding 1939 Occupation and Address

Coding Occupational titles - Machine learning was used to classify occupational strings to the Historical International Classification of Occupations (HISCO). Machine learning is an automatic system that is first trained on a set of examples. Each example contains a particular text string together with its correct classification. The classifier builds a predictive model, which is then used to classify unseen strings. To apply this approach to classifying occupational titles, we used the Mahout machine learning framework (Apache Software Foundation 2011) and evaluated a number of different Machine Learning approaches. We are still assessing quality but in the training data we achieved 82.0 ± 1.5% for micro-precision/recall.



Geocoding - For geocoding, we used the 'Historical Address Gecoding – GIS' (HAG-GIS) software package we have developed for geocoding historical data. Based on the Python language using the SQLite database. The geocoding process has two distinct phases: a) the automated phase where the HAG-GIS system preforms an exact and a fuzzy matching to link each record to a particular geographical reference point, and b) the manual phase where the clerks check/edit the unmatched addresses using the historical maps from Ordnance Survey. The matching is based on contemporary roads, with the search for unmatched addresses aided by the identification of a likely area of search through the production of pseudo boundaries from within the data.

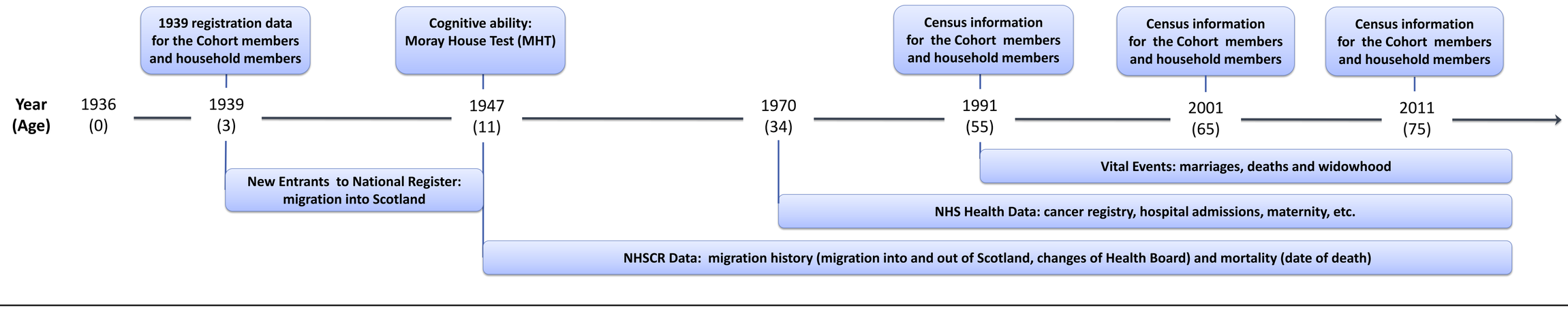
Comments

This paper presents a pragmatic approach of retrospectively creating new birth cohort studies from routine and administrative data. The quality of linkage obtained is extremely good, with a 95.2% backward linkage rate between SMS1947 and 1939 register and a 85.0% forward linkage rate between SMS1947 and SLS. The 96.7% tracing rate at NHSCR is particularly impressive given that the cohort was being traced over six decades later. These high linkage success rates demonstrate that the reliability of the linkage system being used.

Use of the data

- We will use these data to investigate the inter-generational social mobility of this cohort and how their social mobility relates to their cognitive ability and geographic mobility.
- The data will be made available to other researchers via the SLS administration.

Linked data in the 1936 Birth Cohort Study



Acknowledgements

This project is funded by the Medical Research Council as part of a multi-disciplinary programme, Lifelong Health and Wellbeing of the 'Scotland in Miniature' Cohort, led by Professor Ian Deary. 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.