

Record Linkage of Census and Routinely Collected Vital Events Data in the ONS Longitudinal Study

Lin Hattersley, Office for National Statistics, U.K.

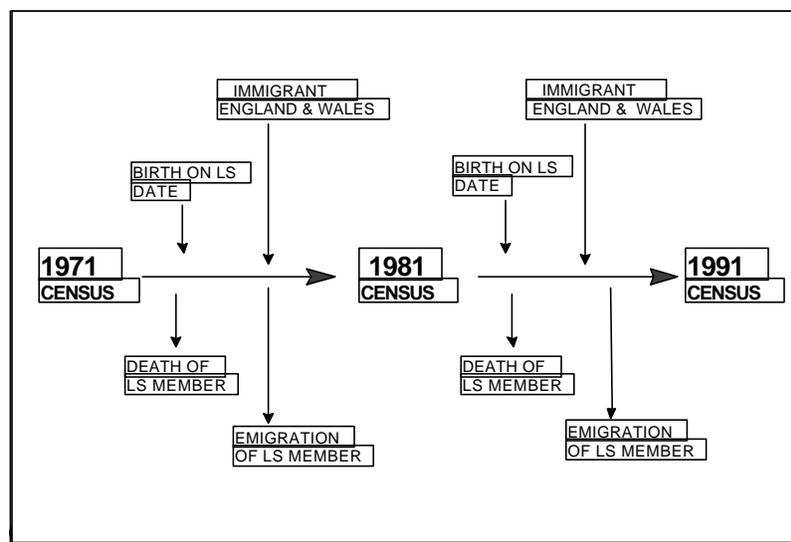
Abstract

Both manual and computerized methods of record linkage are used in the Office for National Statistics' Longitudinal Study (LS) -- a representative one percent sample of the population of England and Wales, containing census and vital events data. Legal restrictions mean that individual name and address data cannot be carried on either census or vital events computer files. Linkage of records has to be achieved by the use of the National Health Central Register (NHSCR) database, where names and addresses are carried together with information on date of birth and medical registration. Once an individual has been identified as a bona-fide LS member and flagged at the NHSCR, data carried on their census record or vital events record(s) can be extracted from the appropriate census file and vital event(s) file and added to the LS database. At no time are the two computer systems linked. This paper will describe the record linkage process and touch on some of the key confidentiality concerns.

What Is the ONS Longitudinal Study?

The ONS Longitudinal Study (LS) is a representative 1 percent sample of the population of England and Wales containing linked census and vital events data. The study was begun in 1974 with a sample drawn from the population enumerated at the 1971 Census using four possible dates of birth in any year as the sampling criterion. Subsequent samples have been drawn and linked from the 1981 and 1991 Censuses using the LS dates of birth. Population change is reflected by the addition of new sample members born on LS dates and the recording of exits via death or emigration. The structure of the population in the LS is shown below.

Figure 1. -- The Structure of the ONS Longitudinal Study



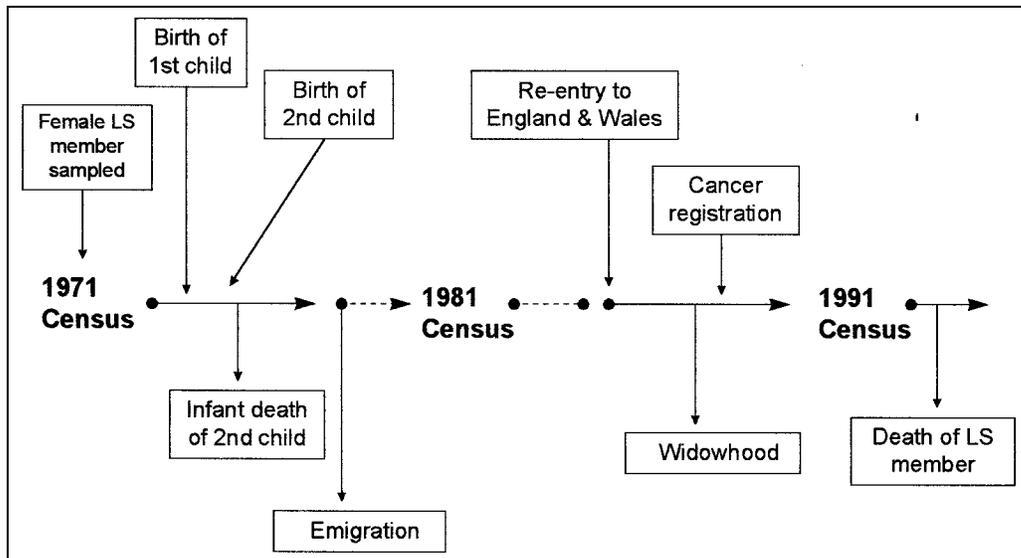
At each of the

er (non-LS) members

in the same household are included in the database. However, it should be noted that linkages of routinely collected events data are only performed for the LS members. The household an LS member resides in at one Census may well be different from the household they are part of in the next, and other (non-LS) household members may therefore change over time.

Routinely collected data on the mortality, fertility, cancer registrations, infant mortality of children born to LS sample mothers, widow(er)hoods and migration of LS members are linked into the sample using the National Health Service Central Register to perform the link (Figure 2). Marriages and divorces cannot be linked to the sample in Britain as the marriage certificate includes age, not date of birth.

Figure 2. -- Event Linkage --The Event History of an LS Member



Creation of the Sample and Methods of Linkage

Linkage methods vary depending on the source of data, but all linkages are made using the National Health Service Central Register (NHSCR). NHSCR performs the vital registration function for England and Wales and is part of the Office for National Statistics. The register was begun in 1939 using the data from the full census of the population carried out on the outbreak of the Second World War. Each enumerated individual was given an identification number which was used to allocate food rationing cards. This number became the National Health Service (NHS) number in 1948 when the NHS was created. Subsequently NHS numbers were issued at birth, or if the person was an immigrant, an NHS number was allocated when they first signed on with a General Practitioner (GP). The NHS number is thus the only identification number that is almost universally held among the population of England and Wales.

NHSCR was computerized in 1991 and prior to that date all records were kept in hand written registers containing one line per person in NHS number order. Events such as births, deaths, cancer registrations, enlistment into the armed forces, entries into long-stay psychiatric hospitals, re-entries to the NHS, embarkation's and internal migration were noted in the registers together with any ciphers denoting membership of medical research studies. In 1991 an electronic register was created.

The Creation of the Original LS Sample

When the LS was begun in 1974 an index card was created for each potential sample member who was

born on an LS date and enumerated in the 1971 Census. A unique 8 digit number was assigned to each LS member and printed on each card together with information that could identify the relevant census forms (such as ward, form number and enumeration district, sex, date of birth, marital status, person number and a usual residence indicator). The relevant census forms were then selected and from these name, usual address and enumeration address were written onto the cards. The cards were then sorted alphabetically and sent to NHSCR where they were matched against the registers. NHS numbers were added to the cards if the person was registered, and the register entries were flagged as LS. These cards were then used to create an LS alphabetical index held at NHSCR.

The essential element in the linkage of events to LS members is the possession of an NHS number and their presence as a member of the NHS register. Those LS members who do not possess NHS numbers are known as “*not traced*” and although Census data can be linked to them, vital event notifications, which are used by NHSCR in maintaining the registers, cannot. By the end of 1976 all but 3.2 percent of the 1971 sample LS members were traced in the register.

Different mechanisms are employed for census record linkage and event record linkage but both are covered by Acts of Parliament which restrict the use of certain data and at present prevent an electronic link being performed between the computer systems of NHSCR and the rest of ONS. Census data is covered by the Census Act which prevents the use of any data that can be used to identify an individual. As a result all completed census forms are stored for one hundred years before public release. Data from the schedules are held in electronic form but exclude names and addresses by law. However, dates of birth of all persons enumerated on each census form are included in the data. This inclusion of date of birth allows the identification of *potential* LS members at any census and together with the data which identifies each form uniquely, allows ONS to extract the forms and provide NHSCR with the names and addresses which can be used to match with their records.

The Linkage of Census Data to the LS

ONS have performed two LS-Census links to date, the first linking the 1971 and 1981 LS Census samples together, the second linking the 1981 and 1991 Census samples. Both LS-Census links were done in the same manner, although the computerisation at NHSCR in 1991 helped to speed the process of the second link.

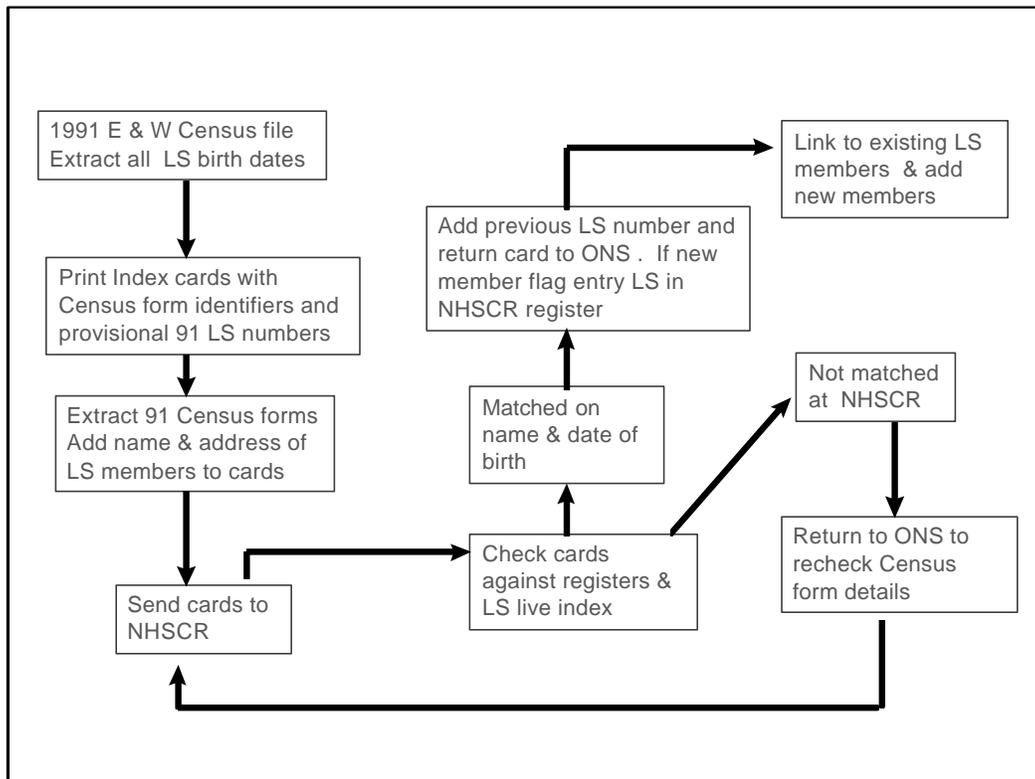
After the 1981 Census, index cards and listings of potential LS members were created from the Census GB Households file by extracting data for each household which contained any person with an LS date of birth. Each potential LS member was allocated a unique 1981 LS serial number which was printed on both the cards and the listings. As in 1971, when the LS was created, the information printed on the index cards was used to locate the Census forms and the name and address were transcribed from the forms. These cards were then sent to NHSCR for matching against the LS alphabetical index. If the LS member already existed in the index (that is had been enumerated in 1971 or had been born or immigrated after the 1971 Census) the 1971 LS number was added to the 1981 card which was then returned to OPCS (now ONS) for processing. If the cards were not matched with any entry in the LS index then a search of the NHS registers was made and if a match was found then the Central Register was flagged LS81 and that person entered the LS as a new member. Further searches against the electoral registers, birth indexes, marriage indexes and the Family Practitioner Committee’s GP patient registers were also made for unmatched cards in 1981. The cards were also checked for “traced” or “not traced” status and were then returned to OPCS for processing as one of five types. These five types were:

- matched to an existing “traced” LS member;
- new “traced” 1981 entrant;

- new “not traced” 1981 entrant;
- matched to an existing “not traced” LS member (these could be “traced” or “not traced” in 1981); or
- matched to a 1971 LS member but a double enumeration.

The LS numbers on the returned cards were validated and the resulting file was run against the 1981 LS Households file in order to add the 1971 LS number or intercensal entry LS number to the records. The final process in the link was the creation of separate LS personal and household files for 1981. After this was completed the cards were returned to NHSCR for addition to the LS index.

Figure 3. -- The Linkage Process



The 1981-1991 LS - Census link was completed in 1995 (Figure 3). Although NHSCR had been computerized in 1991 a manual linkage process was used to fulfill the confidentiality rules. As in 1981 index cards and listings were produced giving census schedule identifiers and 1991 LS serial numbers. The Census forms were extracted and the names and addresses were added for NHSCR identification purposes. Once the cards had been completed and checked they were sent to NHSCR for matching and tracing against the registers and LS indexes. The matching and tracing process was easier and faster than in 1981 as the cards were initially matched and traced against the NHSCR database entries rather than manually against the two rooms full of index cards which formed the LS alphabetical index. Only if no previous LS number existed or there was no entry on the NHSCR database were the cards checked against the clerical registers and indexes. Any cards not matched were returned to OPCS for re-checking against the census forms to identify transcription errors. The NHS number and any pre-1991 LS numbers were added to the cards before their return to OPCS for processing.

How Good Was the Linkage Between Censuses?

The two LS-Census links so far performed have been extremely successful, with at least 90 percent of traced LS member's records being linked together. It should be noted that 97 percent of 1971 LS members, 99 percent of 1981 LS members and 98 percent of 1991 LS members were traced at NHSCR at the time of linkage (Table 1).

Table 1. -- Forward Linkage Rates for the 1971-1981 LS-Census Link and the 1981-1991 LS-Census Link

Forward Linkage Rates					
	1971 Census Sample *	71-81 Linked Sample	1981 Census Sample **	81-91 Linked Sample	1991 Census Sample ***
	N = 512,881		N = 530,248		N = 534,647
Died prior to next census	58,911		58,931		
Embarked prior to next census	5,625		4,399		
Eligible to be in next census	448,345		466,918		
Recorded in next census		408,451		420,267	
Forward linkage rate		91%		90%	

*Traced at NHSCR prior to the 1981 Census (97%).

**Traced at NHSCR prior to the 1991 Census (99%).

***Traced at NHSCR at the 1991 Census (98%).

However, even allowing that LS-Census forward linkage rates were extremely good there were still approximately 10 percent linkage failures at each census. This problem of linkage failure was investigated using the NHSCR records to examine 1 percent samples of linkage failures as part of each of the LS-Census Link exercises.

Table 2. -- Reasons for Failure to Link

	Number Believed to Still be in Sample But Not Found at Census
--	---

Reasons for Failure to Link			
	1971-81 Link N = 39,616	1981-91 Link N = 46,652	All LS Members Who Failed to Link by the 1991 Census* N = 92,580
Date of birth discrepancy between Census & NHSCR	37%	21%	18%
Cancelled NHS registration -- whereabouts not known	6%	9%	16%
Missed event (emigration, death, enlistment)	14%	5%	4%
Not known	10%	5%	18%
Currently registered at NHSCR but not enumerated	38%	61%	44%

*Includes LS members lost to link in 1981 and still not linked in 1991 and LS members linked in 1981 but not in 1991. Excludes LS members who were lost to link in 1981 but were linked in 1991.

The total number of LS members lost to link between 1971 and 1991 was 92,580 (Table 2). Date of birth discrepancies were a major cause of failure to link providing at least 37 percent of failures in 1971. Those in the “Not known” category may well also have included sample members who had given dates of birth other than LS dates on their Census forms. The rise noted in “Cancelled NHS registrations,” which tend to occur if a person has not been seen by their GP for over two years, suggests that many of the persons in this category may have in fact emigrated but not reported it.

Vital Events Linkage

While the LS-Census links only take place once every ten years, vital events linkage occurs annually for most events and six monthly for some. There are two methods of identifying vital events occurring to LS members – firstly, through routine notification of events to NHSCR, where the LS member is identified by the presence of an LS flag in the register; and secondly, through the annual vital events statistics files compiled by ONS. Some types of event, deaths and cancer registrations are identified using both methods as a cross checking device (Table 3).

Table 3. -- Vital Events and the Methods of Linkage

Event Type Currently Collected	Linked Through Routine Notification	Linked Through Stated Date of Birth
New births into sample		X

Births (live & still) to LS mothers		X
Infant deaths of LS mothers children		X
Widowerhoods		X
Deaths of LS members	X	X
Cancer registrations	X	X
Immigrants into sample	X	
Emigrations	X	
Enlistment into armed forces	X	
Re-entries from emigration and enlistment	X	

How the Linkage Process Works

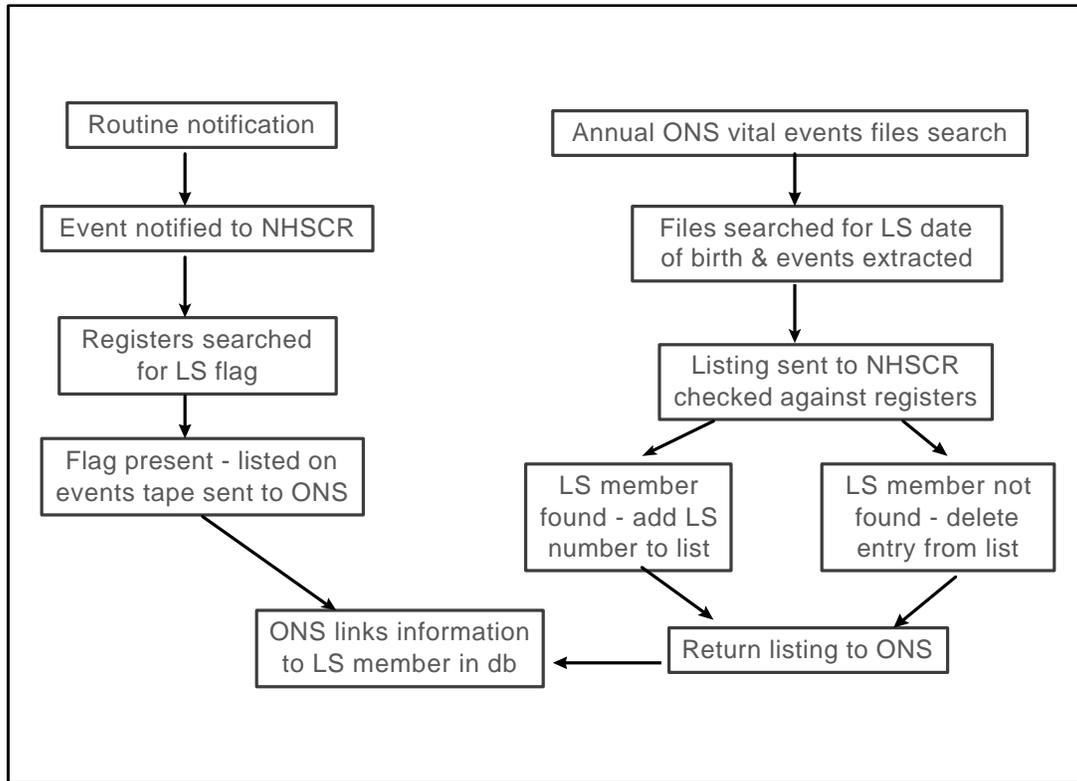
The identification of immigrants into the LS sample, emigrations out of England and Wales by sample members, enlistments into the armed forces and re-entries from emigration or enlistment are all made through the routine notification of these events to NHSCR (Figure 4). When NHSCR updates their database the LS flag is noted for all existing sample members and the details including the LS number are entered onto a tape that is sent twice yearly to ONS to update the LS database. Included in the tapes are details of date of emigration, enlistment or re-entry together with the relevant LS numbers, and for immigrants date of birth and entry details. ONS returns a listing to NHSCR containing the new LS numbers allocated to immigrants joining the sample and this is used by NHSCR to flag their database.

New births into the sample, births to sample mothers, infant deaths of LS members children and widow(er)hoods are all identified using date of birth searches of the annual vital events statistics files. The process involves extracting a subset of data from the statistics files using the LS birth dates as the selection criteria. In the case of new births an LS number is allocated and a listing is sent to NHSCR containing LS number, Date of birth and NHS number. The entry is checked and the LS number is added to the register entry and the new LS member is flagged.

Births to LS mothers are also extracted from the annual England and Wales births file, but the criterion used here is the date of birth of the mother which must be an LS date. A listing including registration details is sent to NHSCR where it is used to extract the relevant birth drafts to identify the name of the mother. The mother's name is then used to find the LS number which is added to the listing which is then returned to ONS for processing.

Infant death details are extracted from the annual deaths file and the mothers date of birth is then matched with the data on the LS births to sample mothers file. Any queries are sent to NHSCR for resolution using the registers. Widow(er)hoods are also linked using the annual deaths file. An LS date of birth search for the surviving spouse is used to extract the data and a listing giving the date of death and registration details is sent to NHSCR. NHSCR have access to the ONS deaths system and use this to identify the names of the deceased and their surviving spouse. The register is then searched for the surviving spouse's name and the LS number extracted and added to the listing.

Figure 4. -- The Linkage Process



How Good Is the Linkage of Events?

The quality of event linkage is extremely good for new births into the sample and deaths occurring to sample members. Virtually 100 percent of these events are linked (Table 4). The rate of linkage for other events is high, with the exception of migration events. Unlike events directly associated with births and deaths which have to be registered within set times by law, migration events do not have to be compulsorily registered. Immigrants can only be linked to the sample when they register with a GP and this may be long after the date of immigration. The date of birth for immigrants is that taken from their NHS registration details and may not be accurate. Certainly, between 1971 and 1981, 62 percent more immigrants were linked to the LS than were expected based on the England and Wales immigration figures. Emigrations of LS members out of England and Wales are only captured if an LS member returns their medical card to their Family Health Service Authority on leaving the country or if the Department of Social Security informs NHSCR when a pensioner or a mother with children is no longer resident. As a result not only are emigrations undercounted but they are often notified to NHSCR many years after the event.

Table 4. -- How Good is the Linkage of Events?

Event	Percentage Linked Between 1971 And 1981 Census	Percentage Linked Between 1981 And 1991 Census
New births into sample	101%	100%

Immigrants into sample	162%	106%
Deaths of sample members	98%	109%
Emigrations of sample members	65%	36%
Births to sample mothers	92%	93%
Widow(er)hoods	77%	84%
Cancer registrations	98%	103%**
Infant mortality	86%*	91%

* Available from 1976

** Available until 1989

Confidentiality Issues

There are two sets of confidentiality issues involved with the maintenance and usage of LS data. First, how to link data without breaching the legal restrictions on the release of census and certain vital statistics data, and second how to ensure that confidentiality is maintained by researchers using the data for analysis.

The processes of data linkage would be accelerated if electronic linkage could be achieved between ONS and NHSCR. However, at present this would contravene all legal requirements including that of current UK data protection legislation. The LS is not a survey where an individual gives their consent for the use of personal data but a study where administrative data collected for other purposes is used to provide a rich source of socio-demographic and mortality data about the England and Wales population over time. Given the restrictions imposed by this situation, the maintenance of the study must not only be done in such a manner as to comply with the legal instruments but must also be publicly seen to do so.

The restrictions on the methods used for linkage of the data also apply to the release of data for analysis by outside researchers. Any data which could conceivably identify an individual such as the LS dates of birth and LS number are used only within the database to achieve linkage between data files. Extraction of data is done within ONS itself and data is only released to researchers in aggregated form which will not permit the identification of an individual.

Conclusion

The LS is a complex linkage study which, by using the only universal identifier held by members of the population of England and Wales (the NHS number), has provided extremely high quality linked data on a 1 percent sample of that population for over 20 years.

The linkage methods used are partially computerised but because of legal restrictions much of the linkage is still labour intensive and reliant on the skills of ONS and NHSCR staff. Automatic linkage would be the ideal, but until it is legally feasible to electronically link the LS system to all other ONS systems (including the Census database) and to NHSCR, this is unlikely to be achieved.