

**STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE**

**STATISTICAL OFFICE OF THE
EUROPEAN COMMUNITIES
(EUROSTAT)**

CONFERENCE OF EUROPEAN STATISTICIANS

Joint ECE/Eurostat Work Session on Registered
Administrative Records for Social and Demographic Statistics
(Geneva, 1 - 3 March 1999)

Session 3, invited paper

The History and Future of Record Linkage in the ONS Longitudinal Study

Prepared by Jillian Smith, Office for National Statistics, United Kingdom

Introduction

1. The Office for National Statistics Longitudinal Study (LS) is a large dynamic record linkage database, with regular updates of information from censuses and vital events, which is used for analysis on a wide range of subjects. It is fully supported for access within a 'safe setting' environment at ONS in London. During its twenty-five year life the study has used a variety of linkage methods, the choice governed by the prevailing technology and confidentiality issues at the time. This paper looks at the evolution of record linkage in the LS and points to issues of current interest.

Context

2. The ONS Longitudinal Study links census and vital event information. A one percent sample of census records was defined from the 1971 Census by taking all individuals born on four dates across the calendar year. This covered about 500,000 people, or one percent of the population of England and Wales, including those in both private and communal households. Subsequently at each census a sample has been drawn on the same basis and linked at individual level into the LS. The database now includes three census samples: 1971, 1981 and 1991 and plans are in preparation to link the 2001 Census sample. The entire census record for the LS member and all members of that person's household are entered into the LS. This contrasts with other census datasets, which are adjusted, aggregated or otherwise limited for confidentiality reasons. In consequence the LS is protected by security measures to provide a safe setting for its use.

3. Vital event information, drawn from the ONS registration systems, has been linked over the period of the study to the individuals in the sample. Two types of entry events replenish the sample. These are births on the four LS dates and immigrations of people with LS dates of birth. Exit events linked to the study are deaths and emigrations. Records of people who have left the study continue to be held in the dataset. Other major events linked are births and infant deaths registered to women in the study, cancer registrations, widowhoods and widowerhoods. Figure 1 shows the structure of the data.

4. The National Health Service Central Register (NHSCR) is used to provide linkage of events to the LS, which itself does not carry identification information such as name and address. LS member's records are flagged with a unique LS identifier on the NHSCR computer systems and extensive manual card systems also provide a range of tracing possibilities in cases of difficulty.

5. The LS data are maintained in a 32 file database in Model 204. Data are extracted from the database into subsets of information tailored for particular analyses which are completed on site at ONS. Users are able to receive their work in several forms, ranging from summary datasets for further analysis to final analyses from a range of software packages.

The Structure of the ONS Longitudinal Study (LS) and events recorded from 1971-1995**

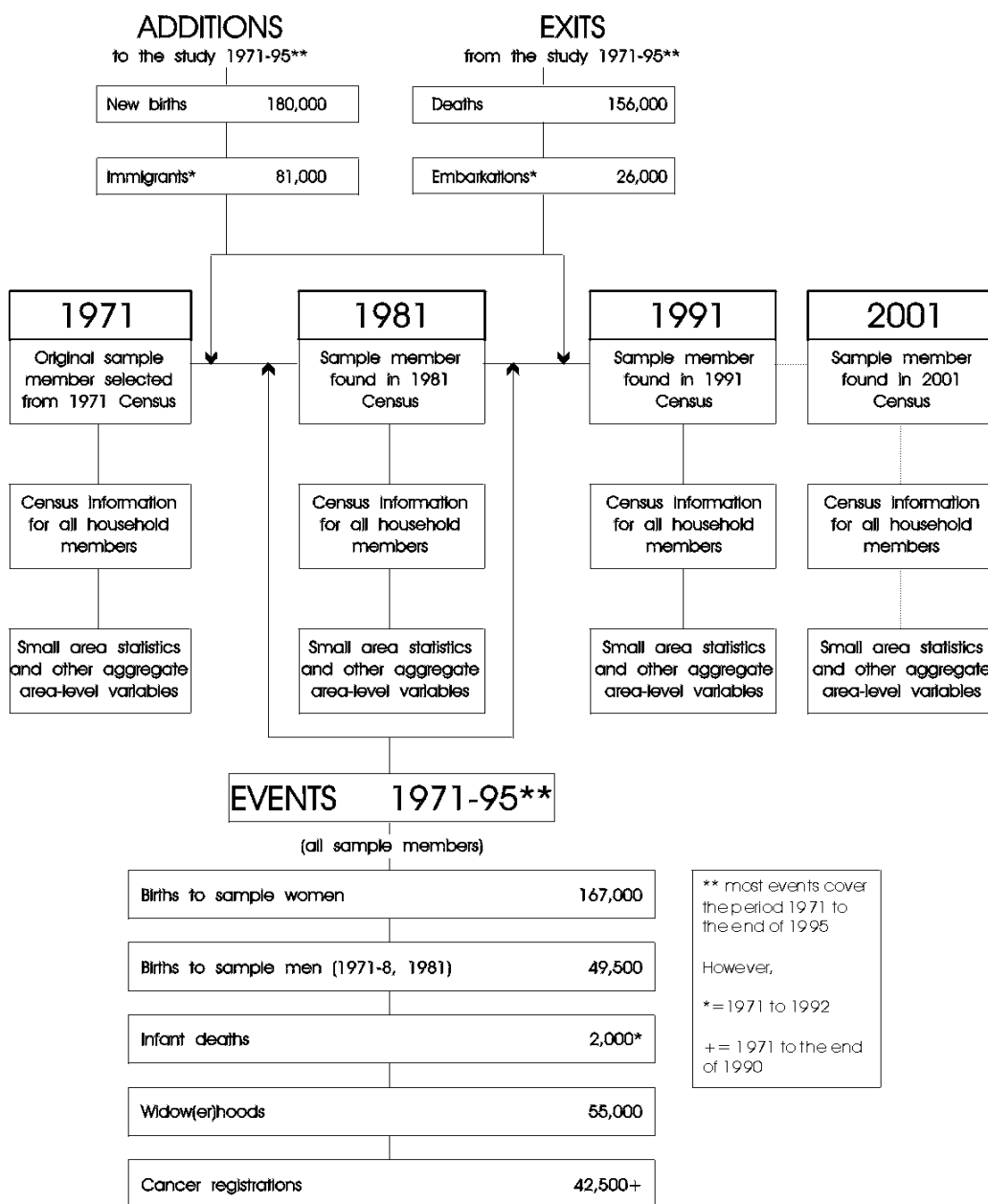


Figure 1

6. An important feature of the LS is the provision of facilities for supported research use within the safe setting. This is both a requirement if the confidential data is to be used and an asset to the end user through provision of skilled analysis expertise to assist work with this complex data. The result is a developed and supported research environment in terms of computer facilities, documentation, quality information and practical data preparation and research skills. The support work is shared through a partnership between the ONS and the Economic and Social Research Council (ESRC) who together fund staff at the Centre for Longitudinal Studies (CLS) at the Institute of Education, London University and within ONS itself. These staff are engaged on supporting analysis for a wide range of users of the LS, from academics to government departments.

7. Researchers are asked to submit applications to an LS research board, giving details which allow an assessment of the scale of work and confidentiality implications of their projects. Each project is allocated a support person who will help with applications, liaise throughout the project and, in many cases, complete the hands-on analysis. Outputs are cleared for publication or release by ONS and CLS staff to ensure confidentiality standards are maintained.

8. The LS is among the largest longitudinal datasets in the country. This allows the study of small groups, change over time, geographic variations, migration and mobility. The study complements other datasets, which may be smaller, and more specialised (eg. the British Birth Cohort Studies) or, using different design criteria, cover shorter periods in more detail (eg. the British Household Panel Study). When the studies are used together, these complementarities help to alleviate deficiencies in each, such as, in the LS, the 10 year gap between censuses. On the other hand the regularity of the LS census points and the long time span allows the study to provide checks on representativeness at different time points to other studies. The study now has 25 years of data. This allows the analysis of intergenerational, demographic and social change over time.

9. The LS is collected with no direct respondent burden. It draws only upon data sets initially collected for other purposes, many required by law.

10. While the LS is unique in the United Kingdom in providing linked census and life event data, some other countries have, or can collate, similar information. This enables collaborative international analysis (Fox , 1989).

Setting up the 1971 Sample

11. In 1974 the initial LS sample was created. The mechanism for linking all future LS data depended on the National Health Service Central Register (NHSCR) acting as an intermediary linkage register. At the time, and until 1991, NHSCR was a manual system.

12. The 1971 LS sample was drawn from the 1971 Census by date of birth. The information required for linkage (eg date of birth, sex, census reference) was printed onto cards then name and address were annotated, by hand, from the census forms. This card index was sorted into alphabetic order and matched against the NHSCR indexes. Where a match was found the NHSCR registers were flagged as members of the LS. The cards were retained to form the LS card index. By 1976, 96.8% of the LS sample had been traced at NHSCR.

Linking Events

13. Once the LS members were flagged on NHSCR, linkage of events into the LS became possible. Events that are notified to NHSCR trigger the updating of the LS member's records.

Such events are immigration and emigration, deaths and cancer registrations. Tapes of the event details are prepared twice yearly, from the NHSCR computer systems, for updating the LS database.

14. An alternative way of capturing events to LS members is through date of birth searches on ONS's annual registers of births, deaths and cancer registrations. Listings of events for people with LS dates of birth are prepared and sent to NHSCR where they are checked and entered where necessary (eg for new births) on the computer systems. The LS identifiers are then used to add the events to the LS database.

15. In the case of deaths and cancer registrations, linkage is completed both via flags at NHSCR and through date of birth searches on vital event data files. Table 1 summarizes the linkage methods for events.

Table 1: The method of linkage used for each event

Event type	NHSCR flag	Date of birth search
a. New births		*
b. Births (and stillbirths) to LS sample member		*
c. Infant deaths o LS sample members		*
d. Deaths to LS sample members	*	*
e. Immigrants	*	
f. Cancer registrations	*	*
g. Widow(er)hoods		*
h. Emigrations	*	
i. Enlistments into the armed forces	*	
j. Entries into long-stay psychiatric hospital (1971-1983)	*	
k. Re-instatements to NHS from h-j	*	
l. Internal migration	*	

16. Estimates of the quality of event linkage are a vital service in terms of documentation and advice for users of the LS and also for users of other datasets who may use the LS itself as a benchmark for comparison. A range of sources is used to generate event rates for the England and Wales area. But gold-standard comparators for some events are non-existent and for others difficult to calculate and deficient in quality. Nevertheless extensive work is carried out to estimate yearly expected event rates by a range of factors such as sex and age. These event rates are adjusted to account for known influences such as the day of the week on which the LS date of birth occurs. The results of this work are fully documented in Hattersley and Creeser (1995). The summary table of event linkage rates given here (table 2) illustrates the particular problems in the areas of immigration and emigration where both the event numbers picked up in the LS and the national estimates are known to be very deficient. This results in linkage rates which range from 162% for immigrants in the first decade to 36% for emigrants in the second decade. This reflects the mismatch of numerators and denominators used to calculate the figures and at least serves as a warning to users.

Table 2: Event Linkage Rates **

Events	Percentage linked between 1971 and 1981 Census, %	Percentage linked between 1981 and 1991 Census, %	Percentage linked from 1991 onwards, %
New Births	101	100	103

Immigrants*	162	106	172
Deaths	98	109	93
Emigrants*	65	36	37
Births to Sample Mothers	92	93	97
Cancers	98	104	101
Widow(er)hoods	77	84	88
Infant mortality	86	91	106

** Number of occurrences in LS divided by number expected in LS (taken from England and Wales estimates), times 100. Linkage figures of over 100 per cent are found when the LS samples more than 1 per cent of the estimated national occurrences of an event.

* Migration data linkage figures are known to be inaccurate because of numerator/denominator mismatch. The denominators (England and Wales migration figures) are estimated from surveys. The numerators (LS migration figures) are supplied from NHSCR notifications. Both are thought to be understatements.

Census Links

17. Every ten years since 1981 a major project is mounted to add the new LS Census sample to the LS database. The LS sample criteria, the four dates of birth, are used to draw this sample from the census. The sample is matched to existing LS records using NHSCR as the intermediary linkage mechanism. Whether or not a match is successfully achieved, the census data for the LS member and that person's household are added into the LS database. A high level of linkage is however vital to the continued quality of the LS and the future capture of event information, so extensive efforts are made to reduce non-matches to a minimum. The detective work made possible by the comprehensive registration systems held by NHSCR allowed forward linkage rates of 90% for each of the 1981 and 1991 Censuses (Table 3). At the time of writing (Dec 1998) plans and tests are in progress for the next LS Census sample to be matched into the database. When this work is complete users will be able to draw upon four census data points and all the intermediary events.

Table 3: LS/Census Forward Linkage Rates, per cent

LS/Census Link	Forward Linkage Rates*
1971-1981	90
1981-1991	90
1971-1991	87

* The forward linkage rate is a measure of successful follow-up of the LS members who are thought to be alive and present in the country at any census: the percentage of the pre-existing LS sample 'found' at the census.

18. Circumstances and technology have meant that each successive link has been completed using progressively automated methods. So the history of census linkage into the LS neatly reflects the improved methods available.

The 1981 LS/Census Link: manual record to manual record

19. The 1981 LS sample was drawn from the census dataset by again producing a full set of cards and hand annotating name and address. These cards were sorted into alphabetic order, then matched first against the LS card index files. Those not found by this method were then searched for by name in the NHSCR alphabetic card indexes and, if found, flagged as newly traced LS

members. Finally those names not found in NHSCR were entered into the LS database as new but un-traced members.

The 1991 Link: manual record to computer record

20. In 1991 the same linking process was followed except that, since NHSCR was now computerised, the LS cards could be matched first against the computer database rather than the LS card index files. This saved time in two ways. Firstly there was no need to sort the LS cards into any order; secondly the NHSCR computer system could be used to call up a list of possible matches when identification details were typed in. Using this process 95% of the 1991 LS cards were linked to a previously flagged LS record and 3% were entered as new LS members.

New Linkage Methods for the LS Census Link in 2001

21. In preparation for the 2001 Census LS link, the 1997 Census Test data has been used to try out new ways of linking the LS sample. The Census Test was carried out in specific areas of the country to test possible census questions and methods. The resulting computer readable file of about 100,000 records, which included names for the first time, was automatically matched against the NHSCR database using five possible tracing routes (fig 2). To achieve the largest possible sample all the test records were used at this stage, regardless of LS dates of birth.

Figure 2: Tracing Routes

1. First 3 characters of surname + First character of forename + Full date of birth
2. First 3 characters of surname + Other initial + Full date of birth
3. Full surname + Forename + 2 out of 3 parts of date of birth
4. Date of birth + sex + FHSA posting*
5. Surname + forename + FHSA posting*

* FHSA posting is health area code applied when a person registers with GP.

22. This achieved a match rate of 70%, which was raised to 89% after additional manual searches. Further matching work, using only the LS element of the sample (1,292 records), gave a final match rate of 96%.

23. The margins of error in translating this test into resource estimates for the full census are wide, especially since the test areas were not representative of the full population but were chosen for particular characteristics which would test out new census methods and included areas which are difficult to enumerate (eg inner city areas). However, it is already obvious that a 70% automatic match will greatly reduce the resources needed for the link next time. Further tests are planned before linkage of the 2001 Census data takes place.

Current Issues

Reliance on Record Systems

24. The success of the LS is entirely reliant upon the excellent manual, and since 1991, computer systems at NHSCR. These systems, are maintained at a high level of accuracy in order to supply information for the per capita payment of National Health Service general practitioners (doctors). As a by-product they have provided up-to-date linkage information which is needed to maintain the LS. Without such a tracking system no longitudinal dataset, like the LS, can survive and thrive as a dynamically maintained, representative resource. So most longitudinal surveys with a constant sample have panel maintenance systems, for example the USA's Panel Study of Income Dynamics (PSID) and the British Household Panel Study (BHPS). Other administrative record linkage studies tend to rely on unique identifiers as found in the USA, Nordic countries and Israel or have needed to invest heavily in constructing their own registry system, as done for France's Échantillon Démographique Permanent (EDP). The importance and reliance on the quality of NHSCR for the LS can hardly be overemphasised.

Confidentiality

25. Maintaining confidentiality and public confidence in that confidentiality is the primary concern of those who manage the LS. Use of this powerful dataset can only be allowed if the anonymity of the sample is preserved. As a result of this, changes of linkage methods are guided and tempered by their wider acceptability. What is made possible through technology tends to run ahead of acceptability and will probably always do so. The limitation this imposes on the data and its use can be reconciled when the alternative is no LS at all.

Conclusion

26. It is as well for the continuation and survival of longitudinal datasets that technology has and still does progress to provide ever more efficient ways of maintaining them. Longitudinal data have never been cheap to produce, but computer matching techniques have greatly reduced those costs. At the same time this progressive potential for linkage has been accompanied by concern for the protection of the information produced. Each step towards quicker, easier linkage must be matched with vigilance in the maintenance of confidentiality. The ONS Longitudinal Study now has a long history of development in both linkage and confidentiality techniques.

References

Fox, J. (ed.) (1989), *Health Inequalities in European Countries*, Gower.

Hattersley, L. and Creeser, R. (1995), *Longitudinal Study 1971-1991: History, organisation and quality of data*, HMSO, London.
