THE SCOTTISH LONGITUDINAL STUDY
An Introduction

LSCS Working Paper 1.0

16 April 2007

Lin Hattersley
LSCS & General Register Office for Scotland

Paul Boyle
LSCS & University of St Andrews

Census.ac.uk
Moving you closer to the data

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

Contents

1 Introduction

The *Longitudinal Studies Centre – Scotland* (LSCS) is an ESRC-funded centre of excellence for the creation, manipulation and analysis of longitudinal data. Its main project has been the establishment of the *Scottish Longitudinal Study* (SLS) which is a large scale linkage study created from the linkage of data from routine administrative and statistical sources. These include Census data, vital events data (births, deaths, marriages), National Health Service Central Register (NHSCR) data (migration in or out of Scotland) and NHS data (cancer registrations and hospital discharges).

A Longitudinal Study (LS) was established in England and Wales in 1971 and it now contains over 30 years of data for 1% of the English and Welsh population (Hattersley and Creeser 1995). This study was set up within the Office of Population Censuses and Surveys, now the Office for National Statistics (ONS), and is funded by ONS and the Economic and Social Research Council (ESRC). The LS has been used to look at a wide range of research questions including occupational mortality, fertility changes, family reconstitution, women's occupations, inequalities in health, ethnic health, migration patterns etc. (see the LS publications list at http://www.celsius.lshtm.ac.uk referenced 20/03/07). Much of this work has fed into government social policy. Although Scotland did create a Longitudinal Study at the same time the LS was created for England and Wales, a 1% sample (around 50,000 people) proved to be too small to allow research on many of the epidemiological and socio-demographic questions of importance to Scotland. The original Scottish study was discontinued in 1981 and the data were not retained.

1.1 Why was the SLS set up?

Scotland is singularly disadvantaged relative to England in the poverty of its longitudinal databases (i.e. databases that *link individuals' characteristics through time*, allowing changing circumstances to be investigated) on demographic, socio-economic and health information. This is especially problematic because Scotland is quite distinct from England and Wales in numerous ways. For example, mortality rates are higher, fertility rates are lower, the population is ageing faster and more people live in deprived circumstances than in England and Wales (Boyle and Graham 2003, GROS 2006, Swerdlow *et al*. 1998). Thus, there were a range of

social research questions which needed to be answered, which were being ignored because of the lack of suitable secondary datasets.

From a practical point of view, establishing a Scottish Longitudinal Study had become more feasible than in the past. With improvements in computing data linkage is now easier and the quality of the electronically held datasets that the SLS draws upon is very good. Also, there is a growing number of academic and government researchers who recognise the value of longitudinal data to answer a range of complex research questions.

1.2 What is the SLS?

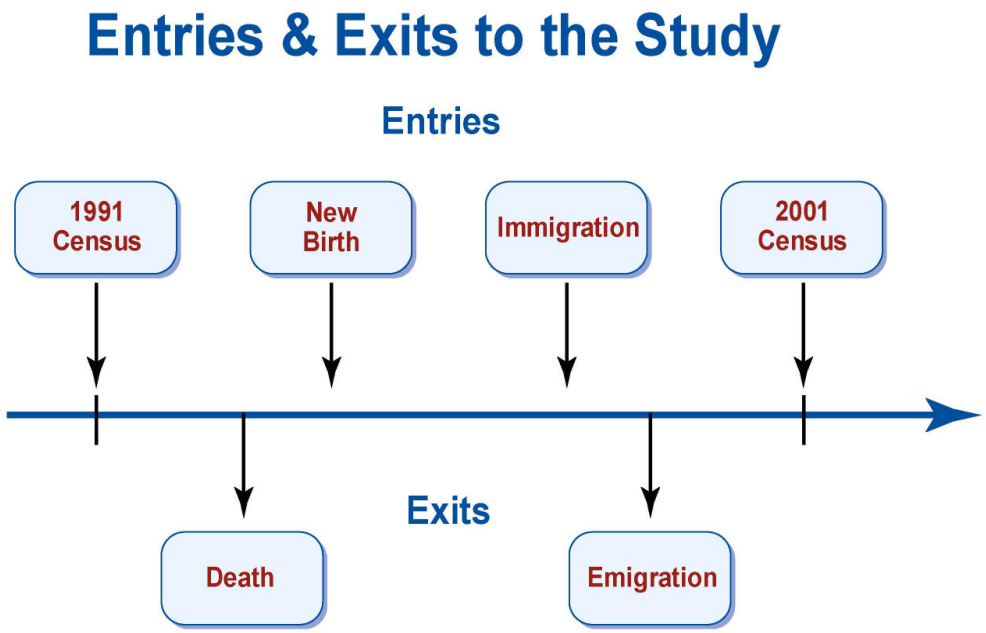
Longitudinal studies have a long history in British social and epidemiological research but most are based on surveys or panel studies (http://www.esds.ac.uk/longitudinal referenced 20/03/07). Some studies use sampling with replacement but most do not. However, all these studies rely on re-interviews of the same persons over time and a high proportion of study members become lost to follow-up.

The SLS, like the LS before it, has been set up to collect data that is either required by law (Census, birth registration, death registration, marriage registration) or is a standard administrative function within Britain. As a result attrition rates are extremely low and linkage rates for events tend to be very high. The SLS is designed to provide a 5.5% representative sample of the Scottish population starting with data drawn from the 1991 Census. The sample is selected using 20 semi-random dates of birth occurring in any year. The SLS date distribution follows the annual Scottish births distribution pattern and included within these dates are the four England and Wales LS dates. The inclusion of the LS dates within the SLS provides the opportunity for the construction of a 1% UK longitudinal study dataset in the future (England, Wales, Northern Ireland and Scotland).

The SLS is a dynamic dataset which is continually updated. Initially, data were extracted for all respondents in the 1991 Census with one of the correct birthdates. In addition, any persons born on SLS dates are included in the sample if they were either born in Scotland or have immigrated into Scotland from elsewhere between the 1991 and 2001 Censuses. Further information is collected about study members in the 2001 Census, when some new members are identified (for example, some

people will move into Scotland between the censuses, but may take some time before they register with a General Practitioner – as described below, it is through the registration with a GP that new immigrants are identified). Exits from the study are either by death or emigration and re-entries can occur into the study if emigrants return to Scotland (Figure 1).

Figure 1 Entries to and exits from the SLS



Events collected for the SLS members include births of new SLS members into the study, immigrations of new members into the study, births, stillbirths and infant mortality occurring to sample members (where the mother and/or the father is the SLS member), widow(er)hoods (where the SLS member is the surviving spouse), deaths, cancer registrations, hospital discharges, marriages (where the bride and/or groom is the sample member), divorces (where the husband and/or wife is the sample member)[1], emigrations out of Scotland and re-entries after earlier emigrations. Table 1 provides figures on the numbers of people captured in the census and vital events data. It should be noted that cancer registration data and hospital discharge data are available, but are not held as part of the SLS database.

[1] The information on divorces has yet to have been added to the sample. It is expected that this will become available shortly.

Instead, they are linked as required for a particular study and, hence, figures are not available for them. It is hoped that Scottish education data (the school census and the SQA exam results) will also be added shortly.

Table 1: Data held in the SLS database as at 16th April 2007 *

Data Source	Number of records
1991 Census - all SLS members selected	270,385
- of whom traced & flagged at NHSCR	265,321
Entry events 1991 – 2001	
New births of SLS members	32,896
Immigrants born on SLS dates	10,594
Exit events 1991 – 2001	
Deaths	28,898
Embarkations	1,237
Events occurring to SLS members 1991 – 2001	
Births to sample mothers	26,509
Births to sample fathers	23,464
Widow(er)hoods of sample members	10,344
Marriages of sample members	21,146
Infant mortality of children of sample members	266
Stillbirths too sample members	252
2001 Census – all SLS members selected	265,104
- of whom traced & flagged or linked at NHSCR	256,379

* Note some of these numbers may be amended at a later date.

2 The methodology

2.1 Selection of the SLS sample

The SLS is a semi-random population sample based on 20 birth dates occurring in any year. It has been set up to be a nationally representative population sample for Scotland which, unlike some panel studies, will reflect population change within the country over time. Thus, the study is a sample with replacement – that is, as people leave the sample over time by death and emigration out of Scotland others enter by birth and immigration into the country. However, the representativeness of the sample over time will be affected by the accuracy of migration reporting between

censuses. Information is extracted from each subsequent census for those with the correct SLS birth dates. There is a loss of some members to the study at each Census, but also the addition of new members who have not been identified before. These changes reflect emigrations out of Scotland which have not been reported to the National Health Service Central Register (some people will move into Scotland and then take some time before they register with a General Practitioner – see below), immigrations into Scotland where the immigrant has not yet registered with a GP, birth dates being given as an SLS date at one census but not the next, and *vice versa*, and SLS members not completing census forms.

The size of the sample (initially 274,055 persons at the 1991 Census but reduced to 270, 385 after the exclusion of dummy and duplicate forms[2] giving a 5.3% sample of the estimated entire population) has been set to allow the investigation of reasonably rare events.

2.2 The role of the National Health Service Central Register (NHSCR)

The National Health Service Central Register (NHSCR) was an offshoot of the formation of the National Health Service (NHS) in 1948. At the time the NHS was founded no central index of patients who were registered with General Practitioners (GPs) was available. To ensure that the payment of GPs was based on an accurate count of patients a national population index was required. This had been created at the beginning of the Second World War (September 1939) with an enumeration of the population to create a 'National Register'. This register was used to issue identity cards and ration books, to identify children who would be eligible for evacuation from cities deemed major bombing targets and to identify adults eligible for conscription into the Armed Forces. The NHSCR adopted the National Register and the civil registration number was adopted as the NHS number. National Registration was not abolished until 1952.

The NHSCR database contains the following information for each patient registered with the NHS in Scotland:

[2] Dummy forms were created for people thought to be missing at the time of the Census, which helps provide a more accurate estimate of the entire population. However, it would not be possible to link these 'invented' people with any other source. Duplicates are people captured on more than one Census form; children who spend time with separated parents are a common example.

- NHS Number
- Surname
- Forename
- Sex
- Date of birth
- The Health Board Area of GP registration in Scotland

It is the only database that includes practically all members of the British population from birth. Registration with a doctor originally required that a person be included on the NHS register so that doctors could be paid a per capita fee for each patient. This system of payment was phased out in 2004 but the register of all NHS patients is maintained for administrative reasons. The main purpose of the NHSCR is to ensure that medical records are moved correctly across UK borders. In Scotland it is also used for authenticating the Community Health Index (CHI). There is no other 'complete' population register in Britain. A team of experts work with the NHSCR to provide flagging and tracing for medical research studies and this expertise is being used by the SLS.

All SLS members are flagged in the NHSCR with a unique study number which allows the linkage of data to be done without compromising confidentiality. This process is described below.

2.3 Tracing, flagging and linkage: a step by step guide

Tracing the SLS members, flagging and linking data to them is a major part of the SLS study. This is achieved through the use of the NHSCR database. Here we explain the steps for creating the SLS.

Step 1: Identifying the sample

First, the 20 birthdates were chosen to provide a 5.5% sample of the population. These are spread across the year.

Step 2: Finding those with the correct birthdays in the 1991 census

Our initial sample was drawn from the 1991 census. Much of the data on individuals is held electronically, including each person's date of birth (name is not held). Those with the correct date of birth were extracted.

Step 3: Locating the SLS sample census forms

There were two reasons for returning to the original census forms. First, we needed to know the names of our sample, so that we could link information reliably from other sources (names are not held electronically as there was a guarantee on the 1991 census form that "Name and addresses will not be put into the computer"). Second, because 'difficult to code' census information (e.g. occupation) was only captured electronically for 10% of the population, we needed to transcribe these data for the remainder of our sample (in fact we transcribed the information for our entire sample, to allow us to check our results against the original transcribing for 10% of the cases). Historical paper census forms are archived according to geography by Output Area, within regions. It was therefore possible to identify the box of forms that each of our sample individuals was located in and these forms were extracted by hand.

Step 4: Flagging the SLS sample in the NHSCR database

The name and some basic demographic information about each SLS member was hand written and passed to the NHSCR (we do not retain information on name or address in any form on the SLS database) where they were identified and 'flagged'. If the NHSCR could not find the person from the information we provided then further demographic information about the other household members was requested. Approximately 10% of records for SLS potential members needed extra information for tracing. Once the person is flagged on this 'third party' dataset and given an SLS number, it allows us to link other data to the SLS, without us ever needing to see names or address information. There are a small proportion of SLS members who enter at Census but are not traced. Many of these are persons such as USAAF personnel who have completed the Census form and have one of the SLS birth dates but will never be registered with the National Health Service. Some of the 'not traced' SLS members are immigrants into Scotland who have not yet registered with a General Practitioner at the time of the Census. These persons will be traced when they finally register with a GP and get an entry in the NHS register. Others not traced are the result of entries on the Census form where an SLS date of birth has been

given but that person does not have an entry with the same date of birth on the NHSCR register. As an arbitrary decision was taken to include as members all those persons who had an SLS date of birth given on their entry record these persons will never be traced. The final tracing rate of 98% was impressive, since we were attempting to trace people based on their 1991 Census details – many will have moved home since then, and a considerable number of women changed name through marriage. A total of 3,670 of the potential members were found to be either duplicate entries or dummies and were removed from the 1991 sample leaving a final total of 270,385 SLS members or 5.3% of the entire estimated 1991 Census population[3].

Step 5: Linking vital events data to a SLS person using the NHSCR

The NHSCR is used as a 'Chinese Wall' where other data providers give the NHSCR information on the people in their dataset with the correct birthdates (e.g. information on those who have died in a particular year, extracted from the vital events data), NHSCR identify them on their dataset and then pass the SLS number back to them. The minimum set of data required by NHSCR to find a person in their database is forename, surname, date of birth and sex. The electronic method of linkage used by the NHSCR team is known as 'exact linkage' as it uses identifiers held on the NHSCR database for a person that must match exactly with the set run against it. Any records not matched automatically were queried and resolved manually. Once the sample provided from the data provider has been identified the SLS number is added and passed back to the data provider. Thus, the data provider now holds data including vital events information, name and address and SLS number for those with correct SLS dates of birth. The name and address information is then stripped from the vital events information before it is passed to the SLS team for linkage through the unique SLS number[4].

Step 6: The process for the 2001 census

[3] See LSCS Working Paper *The Scottish Longitudinal Study: A technical guide to the creation and quality of the 1991 Census SLS sample*

[4] See LSCS Working Paper *The Scottish Longitudinal Study: A technical guide to the quality of vital events data 1991-2001*

When the 2001 SLS Census sample was drawn it required both flagging (for new members who had not been found before that Census) and linkage of data for existing members. Unlike the 1991 Census, in 2001 the names on the census forms were allowed to be computerized but this was done using optical scanning methods which caused some problems in interpretation. However, it did make it easier to provide files for flagging and linkage automatically. As a result, unlike 1991 where only exact matching (auto-matched against the NHSCR database) and manual matching methods were used, it was decided to use a three-stage process. First, an exact match was undertaken, followed by a probability match against those cases that had not been found at the exact match stage. Only those potential SLS members who were still not found after probability matching were then searched for manually. There were 268,428 potential SLS members that were found initially of whom 98% proved to be actual SLS members[5].

Step 7: Linkage of cancer and hospital admissions data

Because health data are particularly dynamic they are not held as part of the SLS database. Data from ISD are only drawn when requested for a specific study. A different method therefore had to be found to allow the linkage of cancer and hospital admissions data to SLS members. This was done by creating a flagging lookup table to be used with the health datasets controlled by the NHS Information Statistics Directorate (ISD). The flagging was undertaken by ISD and NHSCR and involved the submission of a dataset from NHSCR of all SLS members flagged up to and including the 2001 Census. This dataset consisted of the encrypted SLS number, NHS number, CHI number, surname, forename, sex, date of birth and the Health Board Area of GP registration in Scotland for each SLS member. It was run against the ISD health database and where a match was found the ISD unique patient identifier was pulled out and put into a secure lookup table together with the matching encrypted SLS number. ISD used probabilistic matching methods to achieve the match. The lookup table is updated yearly to include new entries to the SLS through births and immigrations. Linkage of data is achieved by requesting ISD to extract the variables of interest for SLS members given certain parameters such as sex and age. On receipt of the linked health file the SLS numbers are decrypted to allow linkage with other SLS data required in the particular research study.

[5] See LSCS Working Paper *The Scottish Longitudinal Study: A technical guide to the 2001 Census SLS sample*

Thus, a complex system is in place which allows anonymous individual-level data drawn from a range of different sources to be held in the SLS. The process is similar to that used in the ONS Longitudinal Study, although hospital episode data are not included in that study.

2.3 Datasets included in the SLS

The datasets currently included in the SLS database are described in Table 2. Basically, they include Census, vital events and health data. We are currently negotiating access to fertility events data between 1974 and 1991, and school exam results and school census data.

The Census data includes a wide range of demographic, housing, employment and social variables. The vital events data are more limited. Detailed information on the variables held in the SLS is available in the data dictionary which can be found on-line at http://www.lscs.ac.uk/sls/dict.htm (referenced 20/03/07). Information on the health data (cancer registrations and hospital discharges) is not contained in the data dictionary, but can be found in a separate document at the same web address.

Table 2 Data currently held in the Scottish Longitudinal Study

Census

1991 Census data for SLS members including:

- 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

1991 Census data for those living in the same household as a SLS member
Similar information as collected for sample members

2001 Census data for SLS members
Similar data to 1991, but additional information collected in 2001 includes:

- self-rated health
- religion
- caregiving

2001 Census data for those living in the same household as a SLS member
Similar information as collected for sample members.

Vital events

New entry datasets

- New births into the sample
- Immigrants into the sample

Vital events to SLS members

- Births to sample mothers / fathers
- Stillbirths to sample mothers / fathers
- Infant mortality of children of sample mothers / fathers
- Marriages of sample members
- Divorce of sample members
- Deaths of sample members
- Widow(er)hoods of sample members
- Emigration out of Scotland of SLS members
- Re-entries into Scotland after previous emigrations of SLS members

Hospital episodes

Information on inpatients and day cases discharged from NHS hospitals (SMR01)
Information on people admitted to mental illness specialties (SMR04)

Cancer registrations

All cancers for sample members

Note that health events data are only linked as required for approved research studies and are not held as part of the SLS database

2.4 Maintenance and Vital Events updates

The maintenance work on the SLS involves yearly event updates, maintenance of logs and archives and ensuring that back-up systems work. The loading of vital events (births, deaths, marriages) into the SLS database takes place once a year. This equates to 16 files:

- Births
    - 1 birth detail file

    - 3 birth link files (one for births to SLS mothers, one for births to SLS fathers and one for the birth of new SLS members).
- Deaths
    - 1 death detail file
    - 2 death link files (one for deaths of an SLS members and one for the widow(er)hood of a surviving SLS member)
- Marriages
    - 1 marriage detail file
    - 2 marriage link files (one for marriages of SLS brides and one for marriages of SLS grooms)
- Infant mortality
    - 1 infant mortality detail file
    - 2 infant mortality link files (one for infant deaths occurring children of SLS mothers and one for infant deaths occurring to children of SLS fathers)
- Stillbirths
    - 1 stillbirth detail file
    - 2 stillbirth link files (one for stillbirths occurring to SLS mothers and one for stillbirths occurring to SLS fathers).

As each file is received by the SLS Unit frequencies are run for every field in the file. These frequency reports are used to spot major problems early on, before the files are loaded into the SLS database.

For a given year all events for each event type are processed together. The order of event processing is important as entries into the SLS for a particular year must be processed before exits. The order of event processing is therefore:

- Entries to the SLS database are processed first:
    - New Births
    - Immigrants
- Exits from the SLS database are processed next:
    - Deaths
    - Emigrations

- Re-entries of SLS members into Scotland from previous emigrations must be processed after exits
- Other events can then be processed in any order
    - Births
    - Marriages
    - Infant mortality
    - Stillbirths
    - Widow(er)hood

3 Confidentiality

The SLS is a large and complex dataset which includes a range of sensitive personal information about individuals. It is essential that people's privacy is protected and that confidentiality is maintained and we therefore use a series of measures to ensure this.

First, strict controls are in place in relation to the dataset itself. The SLS is based on individual-level data for a sample of 20 birthdates. Only a small group of researchers who are responsible for maintaining the dataset are aware of these dates. The dataset is anonymous as neither name nor address is included and, as described above, the method of flagging and linking data involves a complex system designed purely to maintain anonymity in the SLS database.

Second, the environment in which the data are managed is strictly supervised. The dataset is held in a GROS building where staff are required to wear passes at all times. Visitors wear easily identifiable passes and are required to be escorted in the building at all times. The data sit on a stand-alone network which is password protected. Access to this computer is only possible in two rooms, both of which are keypad protected.

Third, the creation, maintenance and use of the SLS is overseen by a Steering Committee and every proposed project is considered by the SLS Research Board which grants permission for studies to be undertaken. No projects are permitted which might allow individuals to be identified.

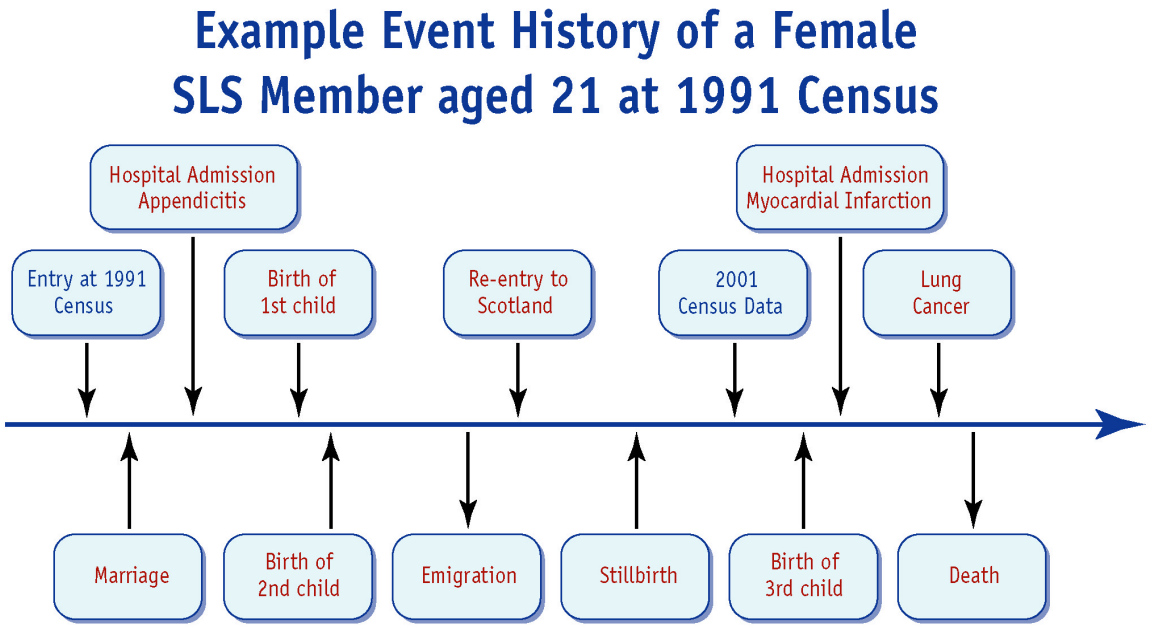
Fourth, once a project has been agreed, access to the data is strictly controlled. Unlike many other secondary datasets, we do not make the data publicly available through any of the academic data archives and individual-level raw data are not released. Instead, a subset of the data is created for each project and the researcher can analyse it using two different strategies. Remote access allows the user to send syntax (SPSS, SAS, Stata) which is run on the data on their behalf. The results are checked to make sure that the outputs do not contain identifiable information (for example, table cell counts must be 3 or more) and are then returned to the user. Alternatively, the user can visit a safe setting in GROS where they work on the data alongside a member of the SLS Support Team. Again, only non-disclosive results can be taken from the safe setting.

4 Research Opportunities

The SLS dataset is a remarkable resource which can be used to examine a range of academic and policy relevant questions. Figure 2 provides an imaginary SLS member and demonstrates the range of events that can be linked together for individuals. For example, we can compare individual's changing circumstances in the two decennial censuses. We can relate their (changing) characteristics recorded in these censuses to subsequent health, fertility or death outcomes. We can also link vital events outcomes to each other, such as comparing women's fertility histories and their subsequent health. In addition, with the ability to compare data from the SLS with similar data in England and Wales and, in time, Northern Ireland allows different cultural and policy settings to be compared.

The England and Wales LS has been used in around 600 academic publications to date and details of these can be found at the Celsius website (http://www.celsius.lshtm.ac.uk/ referenced 20/03/07).

Figure 2 An imaginary SLS member



5 Who funds the SLS?

Funding for the establishment of the Longitudinal Studies Centre – Scotland and the creation of the SLS was originally secured from the Scottish Higher Education Funding Council (SHEFC), now the Scottish Funding Council (SFC). On the basis of this funding, a nationally representative 2% sample of the Scottish population was planned. However, significant additional funding was secured from the Chief Scientist's Office (CSO) which enabled us to increase the sample to 5.5%. During this period funds were also provided from the Scottish Executive and, in particular, the General Register Office for Scotland (GROS) who now fund one of the SLS team.

More recently, funds have been secured from the Economic and Social Research Council (ESRC) to provide an academic support service for the study. The LSCS is now part of the ESRC-funded Census Programme.

6 Conclusion

The SLS is a remarkable resource which will underpin a wide range of academic and non-academic research. The dataset is high quality, large and nationally

representative of Scotland. The methods used to construct the dataset are rigorous and reliable. Unlike other types of longitudinal dataset, such as panel studies attrition is not a serious problem with this dataset, although migrants out of Scotland and the small group of people who are not included on a census form are lost to the study.

The longitudinal nature of the dataset allows people's changing circumstances to be explored through time; it is particularly valuable for event history analyses. Many social science and health research questions cannot be adequately addressed using cross-sectional data and the ability to consider transitions into and out of different states, as well as the opportunity to tease out age, period and cohort effects make this a valuable resource.

The wealth of routine administrative datasets upon which it draws means that the range of questions that can be explored is wide. For example, many studies are likely to be conducted on social and economic mobility between censuses; the relationship between census characteristics and vital events, such as fertility or marriage; and the link between various health outcomes and socio-economic and demographic status.

The free academic support provided for this dataset means that potential users can get expert advice on their proposed projects, and help to undertake them. And the different routes for accessing these data mean that the opportunities for undertaking research projects are flexible.

7 References

Boyle PJ and Graham E 2003 Does Scotland need a population policy? *Holyrood Magazine*

General Register Office for Scotland 2006 Scotland's Population 2005. The Registrar General's Annual Review of demographic trends. 151st Edition Edinburgh: GROS

Graham E and Boyle PJ 2003 *Low fertility in Scotland: a wider perspective.* In Randall J Scotland's Population 2002 – The Registrar General's Annual Review of Demographic Trends. General Register Office for Scotland: Edinburgh

Hattersley L and Creeser R 1995 Longitudinal Study 1971 – 1991: History, Organization and Quality of Data, Longitudinal Study Series no. 7 London: HMSO

Swerdlow AJ *et al*, 1998 Trends in cancer incidence and mortality in Scotland: description and possible explanations *British Journal of Cancer* 77 1-54