Chapter 9

Longitudinal Studies in the UK

Chris Dibben, Ian Shuttleworth, Nicola Shelton, Oliver Duke-Williams

Abstract

There are three longitudinal census studies in the UK: the ONS Longitudinal Study (covering England and Wales), the Scottish Longitudinal Study and the Northern Ireland Longitudinal Study. These have been running for different lengths of time: the ONS LS from 1971, the SLS from 1991 and the NILS from 2001, although the NILS has retrospectively added census data back to 1981. The studies vary in their sample size: the ONS LS 1.1% of the population, the SLS 5.5%, and the NILS approximately 28.5%. The samples are constructed on the basis of individuals' birth dates; these dates are not disclosed, and individuals do not know whether or not they are in the sample. All three studies also include linked administrative data (or permit such data to be linked for specific project extracts), with considerable variation in the range of such data.

The data are detailed and thus potentially disclosive. Access to the data is therefore restricted in a number of ways: all researchers must individually have approved status, and each project must also be approved. Having acquired these approvals there are then close restrictions on the use of the data: access is either within a secure facility to which researchers must travel, or via remote submission of a processing script to a support officer, who can return results only if they satisfy certain criteria including minimum cell counts. Additional release criteria then apply to the publication of any results.

The three studies are separate, and cannot be analysed together as a UK-level data set. However, a number of resources have been developed to aid cross-study analysis, including an integrated data dictionary, and a process by which models can be securely and iteratively run across more than one study.

9.1 Introduction

The Longitudinal Studies (hereafter collectively referred to as 'the LSes') are the most complex of all the outputs from the UK censuses. Whilst introduced at separate times, they are now considered to form a distinct family of data resources. There are three separate studies: the ONS Longitudinal Study (ONS LS), the Scottish Longitudinal Study (SLS) and the Northern Ireland Longitudinal Study (NILS). The ONS LS, which covers England and Wales, was the first of these studies to launch and does not refer to its spatial coverage in its name:, the later SLS and NILS have more clarity in their titles. In order to reduce ambiguity and to aid consistency with the other two studies, the ONS LS is sometimes referred to by users as 'the England and Wales LS', but this informal label will not be used in this chapter.

The three LSes have a largely similar design with individuals linked between censuses, although the linking approach in Northern Ireland is different to that used in England and Wales and in Scotland. The availability of data from different time points allows a number of different types of analysis to be conducted. For example prospective analysis of census characteristics, prospective analysis of event level data (from linked administrative data) and retrospective and prospective analysis between census and event data, as well as international comparisons with equivalent data elsewhere.

The three studies vary considerably in other aspects, such as the sample size, number of census years linked, and amount of other data for individuals that is also linked. Data for different censuses and for different administrative items are held in separate tables. When a user data request is being prepared, records are joined across multiple source tables as required, and a single set of output data records are produced for the user to analyse.

The LSes have provided a set of research data that allow researchers to explore a diverse range of demographic and social issues, using a variety of analytical methods. The longest running study is the ONS LS, and the CeLSIUS website identifies, at the time of writing, over 800 research outputs by users of the data, including journal papers, research reports and conference presentations. Research has covered many areas including, for example, associations between unemployment and mortality; variations in social mobility; lone parenthood;

household structures; trends in fertility and labour market behaviour. Extensive research has taken place on links between inequality and health, with work using the ONS LS informing the Black Report (Black *et al.*, 1980), The Acheson Report (Acheson, 1998) and the Marot Review (Marmot et al., 2010). The LS was also a key dataset used in calculations of trends life expectancy used in the Turner Report (2005) reviewing pension savings in the UK and advising on related policy.

9.2 Definitions

Longitudinal studies (or surveys) are ones that involve repeated observation of individuals over a period of time; in the case of human studies the period of time may extend to many years. A number of possible designs exist for such studies: members might be selected through having a common characteristic such as a birth date (or year), or through experiencing a common event (the cohort of people who left school in a particular year or the cohort of people who have had a particular medical intervention, for example). In the context of UK social studies, there are a number of well-known non-census studies, including birth cohort studies (from the 1946 National Survey of Health & Development (Wadsworth et al. 2006), the Millennium Cohort Study (Plewis, 2004)), and. most recently, studies such as Understanding Society (Buck and McFall, 2011) based on a sample of households.

Census LSes are specifically built around samples drawn from the census, but they also contain additional linked data including life events, and health-related and other administrative data. The LSes could therefore be seen as ‘just’ a series of cross-sectional observations that couple detailed demographic data with life events. However, such a view of the LSes would be to miss their most important characteristics: the fact that individuals can be observed at multiple time points across the life course allow the researcher to identify associations between past experience (housing, education, employment etc.) and later life outcomes. This permits richer analysis of cause of death, for example, than would be possible solely using the individual level data recorded on the death registration. Furthermore, it should be noted that to view LS data solely as cross-sectional is also to miss the point that the Samples of Anonymised Records (SAR), as discussed in Chapter 8, may well be a better resource for cross-sectional analysis of individual level data.

9.3 The studies

The three LSes share broadly similar content. They contain data on sample members, who are selected on the basis of their birth date. These data are captured from administrative sources and from census returns. Given the context of this volume, more focus is placed in this chapter on the census data aspects of the LSes, but it is important to recall that much of their strength lies in the administrative data that are linked to the census records. In terms of census data, the three studies share common benefits: that data are captured for sample members and also for 'non-members'.

'Non-members' are other residents in members' households, as identified in each census. Census data are retained for the non-members in a similar fashion to the data for members – that is, with similar coding used for variables. As a sample member ages, so the nature of the associated non-members is likely to change: for a sample member who is a child, the non-members usually consist of their parent/s and possible siblings; by the time that same sample member has become an adult, the associated non-members are more likely to include a partner and the sample members' own children. Whilst LS members are explicitly linked between censuses, this is not the case for non-members; thus, for a hypothetical adult sample member, it is not possible to definitively state that the spouse observed as a non-member in the 2011 Census is in fact the same person as the spouse observed as a non-member in the 2001 Census. Of course, it is possible to draw inferences from other available data about whether or not this is the case; for example it can be identified whether a spousal LS non-member in one census has the same date of birth as in a previous census. The 2001 and 2011 Censuses contained detailed 'relationship matrix' questions showing the relationships between all household members. Given that sample members are selected on birth date it is feasible that a single household will contain more than one sample member. The likelihood that this will occur varies across the three studies, as they all have different sampling fractions.

Unlike birth cohort and other panel studies, people cannot opt-in or opt-out of the LSes: all persons who are born on one of the LS birth dates are included in the sample automatically. There are a number of ways that people may 'enter' or 'exit' the sample: these are enabled through both census records and administrative data. Migrants entering the UK from overseas will be recognised as sample members based on birth date when they enter into the National Health Service (NHS) administrative system (with variations in practice in the different member countries of the UK). Similarly, babies born on a qualifying date effectively become sample members at birth (once the birth is administratively recorded) – they do not have to 'wait' until the next census to be identified as sample members. People may 'leave' the sample through either death or embarkation (emigration). Whilst death is administratively well documented for almost all people, embarkation is more problematic: it is possible for people to leave the UK without notifying the NHS or other administrative data sources. When people do ‘exit’ from the study, their records are retained so that they can still be used for analysis, and, for those who have emigrated, for future continued usage should that person re-enter the country.

When a new wave of census data is added to an LS, it is necessary for tracing to take place – an administrative process by which census records for an individual are linked to national health records, which then permits linkage to earlier census data for the same person. Not all persons who have qualifying dates of birth can be traced, and it is obvious that no tracing and linking processes will be perfect. Ambiguity can arise from errors in form completion, and it is also possible for multiple census records to exist for a given person (for example, students recorded at both a term-time address and at a parental address, and in the most recent census, persons with both an internet-collected record and a paper-collected record). Where multiple records exist, one preferred record has to be identified.

It is useful to recall that the three LSes are all implemented as independent studies: even though there are common sample-membership birth dates, a person moving from (say) Northern Ireland to England would be seen as leaving the NILS, and as a new entrant to ONS LS. The LS study in the destination country would not 'inherit' their earlier census records from their origin country.

Data from the LSes have the potential to be highly disclosive: even for a single census, the combination of responses to census questions (including place of usual residence to an aggregate level) may be unique, but when responses are linked across multiple censuses then the probability of uniqueness rises as the number of attribute fields in each record grows rapidly. The birth dates which are used to draw the samples are not disclosed, which offers a defence against attempts to identify individuals. It is important that identification is not made, because any such identification would disclose one of the sample birth dates, thus revealing that all persons with that birth date were sample members. Thus considerable emphasis is placed both on security of access to the data and also on responsible use of the data by those who are working on approved projects.

ONS LS England and Wales

The original sample for the ONS LS was drawn in 1974 (OPCS, 1973) from individuals recorded in the 1971 Census. Two concerns were identified, justifying a decision to commence a longitudinal study (Hattersley and Creeser, 1995): firstly, that more information on fertility and birth spacing was required (the 1971 Census had included additional questions on fertility), and secondly that occupational data as recorded on death registrations were not ideal for determination of occupational mortality rates, as changes in occupation over a person's life were not recorded.

Hattersley and Creeser (1995) identified a number of methodological developments that permitted a linkage design to be established with a sample drawn from the 1971 Census. Firstly, the 1971 Census included a question asking for respondents' dates of birth, rather than age. This was the first time (excluding the 1966 Sample Census) that full date of birth had been gathered. Similarly, birth and death registrations had included date of birth from 1969 onwards, permitting potential linkage on a date of birth basis. Finally, general advances in information technology in the 1960s had made such a linkage study feasible.

The sample was drawn by selecting four birth dates, giving a sampling fraction of 4/365, or 1.1% of the population of England and Wales. As with all of the studies, these birth dates are not disclosed. The 1971 sample consisted of around 500,000 people, with a similar number of persons (allowing for overall population growth) being sampled at each subsequent census. Sample members are included in all censuses for which they are present and enumerated. More than 200,000 people have been enumerated in five successive censuses (from 1971 to 2011) (Lynch et al, 2015).

In the transition between any two consecutive censuses, some sample members will be lost to the sample either through death or emigration, whilst others will be added to the sample, through birth or immigration. Thus, any child born with an LS sample birth date will automatically become a sample member; similarly someone entering the country (once they enter in to the NHS registration system) with an LS sample birth date will become a sample member. Successful linking clearly depends on the individual being included in the census data capture, and therefore people may effectively leave or enter the record set through enumeration or failure to be enumerated in the census. Blackwell et al (2003) reported tracing rates from 1971 through to 2001 for the ONS LS. These varied from 98.4% in 1991 to 99.3% in 2001; the tracing rate in 2011 was 98.8% (Lynch et al., 2015)

As well as census data, the ONS LS contains linked data on birth and death registrations of sample members, on live births to sample mothers (and, for some time points, on fatherhood), on immigration and emigration (as observed via NHS registration), on cancer registration and on widow(er)hood (death of a sample member's spouse).

SLS

Although a sample similar to the ONS LS was extracted from the Scottish 1971 Census data, the 1% sample (around 50,000 people) was argued 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, therefore, discontinued in 1981 and, unfortunately, the original data from the 1970s erased. Given that Scotland had, compared to England, relatively few longitudinal databases, combined with a growing recognition that a set of fairly unique demographic and health issues were facing policy makers in the country, (for example, mortality rates higher, fertility rates lower, a population ageing faster and more people living in deprived circumstances than in England and Wales), in the 2000s the idea of a Scottish LS was revisited.

Various factors made the construction of a longitudinal study more feasible in the 2000s. The growing awareness of the value of longitudinal data in answering a range of complex research questions among a number of academic and government researchers led to valuable support for the funding requests. Improvements in computing power, data linkage techniques and the quality of electronically held administrative datasets since the 1970s also meant that embarking on such a linkage study was more technically feasible. On the other hand, attempting to locate and transcribe information from some of the historic census and life events records raised a series of challenges. A group of academics requested funding from the then Scottish Higher Education Funding Council (SHEFC), now the Scottish Funding Council (SFC), to establish the Longitudinal Studies Centre - Scotland (LSCS), which is responsible for the establishment, maintenance and support of the SLS. Because of the problems associated with the 1% sample size, identified when the Scottish component of the LS was abandoned, this funding allowed for a 2% nationally representative sample, based on eight birth dates (four of these matched those used in the ONS LS to allow future comparative studies). Further funding to establish the study was then secured from the Scottish Chief Scientist's Office (CSO), which allowed the sample to be extended to 5.5%, based on 20 birth dates. Funding from the Scottish Executive (now Scottish Government) and, more recently, from the ESRC, has since enabled the establishment of the SLS support team that provides tailored, free support to academic researchers wishing to use the dataset.

The SLS is similar to the ONS LS but routinely links not only to life events data but also to secondary health care and more recently, to education census and outcomes data (since 2007) and prescribing data (since 2009). Life events data collected for the SLS members include: births of new SLS members into the study (those born with one of the 20 birth dates), 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 records, marriages (where the bride and/or groom is the sample member; divorces (where the husband and/or wife is the sample member; note that the information on divorces will become available shortly), emigrations out of Scotland and re-entries after earlier emigrations. These events have been added for the period 1991–2013. It is also planned to include fertility events between 1974 (when the information on life events was first collected electronically) and 1991, allowing the construction of a complete fertility histories for some women in the study.

The health data are provided by the Information and Services Division (ISD) of the Scottish NHS. Unlike the vital events data, which are linked into and held on the SLS database along with the census data, due to the dynamic nature of these health data they are linked on a project-by-project basis. As with the ONS LS, these include cancer registrations, which occur to sample members. Unlike the ONS LS, though, it has also been possible to link hospital episode information, allowing studies of a wide range of morbidity outcomes. Recently, use has also been made of the link within the maternity record between mother and child to produce a new cohort, child of the SLS members (COTS) who can be followed up in the health care data.

Until now the only data on educational experience and attainment of SLS members has been the 1991 and 2001 Census data on educational attainment. In 1991, only tertiary qualifications were noted. In 2001 people reported which qualifications they had attained, ranging from O-grades to degree level, but with no details or indication of when they were attained. To augment this, data were obtained from ScotXed, the agency within the Scottish Government responsible for collecting and coordinating data from schools. These consisted of the following: School Census data for every pupil in Local Authority (LA) funded schools, data on attendance, lateness and exclusion from school for the same pupils and attainment data originally collected by the Scottish Qualifications Authority (SQA) giving details of the results for all SQA accredited qualifications for candidates the school years 2007-8 to 2010-11. The School Census data were obtained for censuses conducted in the September of 2007-2010. Attendance and lateness data were collected for these same school years, although the collection was done at the start of the following year. The SQA data were for qualifications examined in the equivalent school years.

Recently, the SLS has been starting to look backwards in time. For a cohort of study members born in 1936, two additional sets of records have been collected and linked: a cognitive ability test they sat in 1947 (aged 11), and the 1939 National Register for them and their family aged 3. Together with the data collected as part of the main study, this has delivered a 1936 cohort with early life conditions, cognitive ability, school outcomes data, middle life occupational information and detailed information after the age of 55. It is hoped to extend this work as increasing amounts of historic administrative data is made machine readable.

NILS

The Northern Ireland Longitudinal Study (NILS) was established in 2006. It arose out of discussions between the academic community in Northern Ireland and the Northern Ireland Statistics and Research Agency (NISRA) prompted by the existence of LSes in other constituent countries of the United Kingdom, but not Northern Ireland. It differed from both the ONS LS and the SLS in that its structure since it was based on health card registrations rather than the census. Crucially, this data spine provides a link to other health and social care data and regular six-monthly address updates with the possibility also of linking census data for NILS members and members of their households. Equally important is the sample fraction of the NILS. The sample uses 104 birth dates including those used in the SLS (which itself includes those birth dates used in the ONS LS). At about 28.5% this fraction is the largest of all the UK LSes and so, although the absolute number of NILS members at each Census is about 500,000 (a similar number to the ONS LS), it is possible to do finely-grained analyses down to the level of Super Output Areas (SOAs) as small numbers in small areas do not raise disclosure problems.

The NILS started by linking just the 2001 Census but its Census data holdings rapidly expanded. The 2011 Census link was completed by 2013 but this was swiftly followed, with support from the ESRC, by retrospective links to the 1981 and 1991 Censuses in 2014 and 2015 respectively, and since a 2021 Census will be taken in Northern Ireland it is expected by the middle of the next decade that there will be five Censuses linked to the NILS which cover forty years of rapid social, economic and political change. There is, however, more to the NILS than this. Information from the Valuation and Land Agency (VLA) has routinely been linked to the NILS from its inception. This provides data on rateable value and other housing characteristics. Via the health and social care spine there are also routine linkages which provide data on births and deaths of NILS members and also births *to* NILS members. There is the potential also for data linkages to be made on request to explore marriages and widowerhoods. The NILS data framework also supports the Northern Ireland Mortality Study (NIMS). This is a way to access data on 100% of deaths in Northern Ireland. There is a 1991 NIMS linked to the 1991 Census, and also 2001 and 2011 NIMS. Given that this covers all deaths it is possible to deal with detailed causes of mortality once sufficient deaths have built up as time elapses from the base Census. Finally, there is the possibility of linking data not routinely available but held by the health and social care system via the mechanism of Distinct Linkage Projects (DLPs) which provide an ethical and tested way to expand the data available to researchers. Examples of these include the use of prescription and cancer screening data.

Looking forward, the prospect of linking the 2021 Census to the NILS has already been mentioned, and this will extend the already rich research potential of the NILS within its current institutional setting. However, the advent of the Administrative Data Research Network (ADRN) across the UK, and the regional Administrative Data Research Centre for Northern Ireland (ADRC-NI) increases the probability of linking the NILS to other administrative datasets from education, justice and social welfare and thus potentially takes the NILS into new territory. These datasets are of research interest but they very often lack covariates so there is a limit to what can be done using them. The potential to link them to the NILS will offer the chance to do analyses with temporal depth and to consider how an individual's current personal and household circumstances (for example, whether they are on jobless benefits or not) can be understood in a life-cycle framework by relating the present to their situation in 2011, 2001, 1991 and 1981. This will add value to the NILS and to the data to which it might linked

9.4 Data use arrangements

Owing to their disclosive nature, the LSes have much stricter access arrangements than most other census related data sets. The arrangements for each of the three studies are similar but not identical, but they share strict concerns about the risks of breach of confidentiality. The path to preventing any such breach is to adopt a number of inter-related strategies. The data can therefore only be used by approved researchers working on approved projects, and working under specific access conditions.

The LSes, and their associated support units, have been running for a considerable number of years. Consequently, access arrangements have adapted over time as permitted by developments in technology and changes in user expectations, but nevertheless have continued to be guided and (relative to other data resources) limited by the overriding concerns of data security.

A common aspect of all three support units is that the support that is given to researchers working in projects is free at the point of use.

CALLS Hub

The Census and Administrative Data Longitudinal Studies (CALLS) Hub was commissioned by the ESRC alongside the re-commissioning of the three research support units for an initial five-year period from 2012 to 2017. Its stated role is to co-ordinate, harmonise and promote the work of the three LS Research Support Units (CeLSIUS, SLS-DSU and NILS-RSU, described below), with the intention of providing a streamlined experience for users. One of the key purposes of the Hub is to act as an initial point for researchers who are contemplating using one or more of the studies. The Hub is a collaboration between the University of St Andrews, University of Edinburgh, and University College London, though the management group also includes the directors of CeLSIUS, SLS-DSU and NILS-RSU.

The Hub acts to combine information about the studies, and also to provide resources including copies of all relevant census forms, and an integrated data dictionary.

CeLSIUS

The Centre for Longitudinal Study Information and User Support (CeLSIUS) provides support for UK academic, statutory and voluntary sector users of ONS-LS. Additional users are supported directly by ONS. CeLSIUS is an ESRC-funded research support unit. It was based at the London School of Hygiene and Tropical Medicine under the directorship of Professor Emily Grundy during the period 2001-2012, and moved location to University College London (UCL) when re-commissioned for the period 2012-2017, under the directorship of Dr Nicola Shelton. Prior to the establishment of CeLSIUS, support for the ONS-LS was provided through the Social Statistics Research Unit at City University from 1982 and from 1998 at the Centre for Longitudinal Studies, Institute of Education.

In order to use the ONS LS, researchers must follow a number of stages. A Research Proposal Form must be submitted, which details the purpose of the intended research, and an LS Supplementary Form which identifies the specific data items and population which are required for the research to be carried out. The research proposal must name all researchers who will be involved in a project, including those who will not necessarily directly carry out analysis (for example, a PhD supervisor). All named researchers must hold ONS Researcher Accreditation, which involves meeting certain criteria and then making an Accredited Researcher Application; as part of gaining accreditation it also necessary to complete certain training. Upon completion of this training, researchers are asked to sign an Accredited Researcher Declaration. ONS Researcher Accreditation lasts for a period of five years.

Each project will be assigned to a Project Officer who will assist with the application and will support the user during the analysis. In practice, researchers are encouraged to contact a support officer prior to submission of the research proposal who will discuss the project and likely variables required.

Having gained accreditation and had a research proposal approved, a project-specific data extract is prepared, and there are then two ways in which researchers can use the data. Direct access to the data extract is possible by using a terminal in a secure setting. A session must be booked in advance, and most ONS LS researchers use terminals at the ONS London offices in Pimlico. No data or notes may be taken out of the secure environment: results can be subsequently sent to users if disclosure control criteria (such as minimum cell counts) are satisfied. Alternatively, users may remotely submit a script for use with one of the supported software packages; scripts are run by a support officer, who will then return results again subject to them meeting the disclosure control criteria.

SLS-DSU

Use of the SLS is supported by a unit from the University of Edinburgh based in offices within the Scottish National Statistical Agency, National Records of Scotland (NRS) under the directorship of Professor Chris Dibben. The unit was originally set up by Professor Paul Boyle, then at the University, of St Andrews and based there until 2014, when it moved, with Dibben, to Edinburgh.

Using the SLS requires some preparatory steps before a researcher can access data. Prior to contacting the support unit it is recommended that a researcher attends an SLS training session, reads through the information on the unit's website, in particular the data dictionary, and have developed a set of research question. Because SLS data are quite different to other types of social science and health data, it is always helpful to have an early conversation with a support officer who will have had many years of experience using LS data to further scope what may be possible. Researchers wishing to use NHS data in their analyses are required to complete an approved safe researcher training course and all researchers need to acquire approved researcher status (assessed on application by the unit). All research needs to be feasible and robust and therefore requires an application to the SLS Research Board who assess whether it should be supported and may provide some advice on how it could be improved. Final approval is granted after both SLS Research Board and all appropriate ethical board approval is gained.

Because of the sensitive nature of the data, direct access to the SLS is only possible on non-networked computers in a safe-setting in Edinburgh, though the support team are able to run syntax provided remotely. The safe setting computers have standard statistical software such as *SPSS*, *SAS*, *R* and *Stata*. After running analyses (or having them run remotely), output files must be cleared by the SLS team before they can be released. The process for clearing final outputs protects the SLS by reducing the risk of disclosure, ensuring that the study and data are properly described and ensuring that the data have been used appropriately.

NILS-RSU

The arrangements for accessing the NILS for research purposes have much in common with those for the other UK longitudinal studies in their generic features. The NILS has been in operation for ten years and the application process from beginning an application to getting data can be speedy and completed within four months. The process starts with researcher validation – an evaluation of whether the person is a 'proper person' to conduct research – and this is assessed by means of a statement of research experience, membership of learned bodies, and relevant publications. The researcher must also be a 'safe researcher'. This means that they have undertaken training on data security and NILS procedures.

Once these hurdles are overcome research ideas can be submitted to the NILS Research Support Unit (NILS-RSU) where guidance is offered on the feasibility of these, the range of available data, and the applications process. Including advice on completing the application form although researchers can make considerable solo progress using online resources such as the NILS data dictionary and metadata. As might be expected there are standard items that are requested such as project title, abstract and the intellectual context for the work, but there are some features that are not seen in the other LSs since relevance to health and social research must be demonstrated and there is also a requirement to show plans for dissemination especially with regard to policy relevance.

The NILS-RSU website provides example forms to guide researchers. The NILS does not provide all the variables held in the database to researchers but only those that are requested. The application should provide a rationale for the variables to be chosen especially when dealing with sensitive information such as religion which is deemed in the NILS data dictionary to be restricted. The applications are assessed by the Research Approvals Group (RAG). This meets every two months and includes representatives from academia, NISRA, the Public Health Agency, and the Social Care Business Services Organisation. It considers applications using eighteen criteria but with two (a longitudinal element and relevance to health and social care research) being essential. The RAG may simply approve the application or it may return it to the researcher with requests for clarification or suggestions for improvement.

Once the project has been approved the requested data are extracted by the NILS-RSU staff. Users receive large text tables which they must import into their chosen statistical software (*SPSS* and *Stata* are supported) where the data can be labelled and prepared as they wish. Typically, several different data tables are received, for example 2001 Census individual data, 2001 Census household data, 2011 Census individual data and 2011 Census household data) and these are linked by the researcher using the index fields provided by NISRA to create the analytical database. Once this is done, the researcher is free to work on his or her project. Normally this is done in the NILS-RSU (there is no facility to work remotely) although it is possible to submit by email SPSS, STATA and R code which can then be run by NILS-RSU staff.

Outputs can be either intermediate or final. Intermediate outputs may be shared within the project team amongst those who are signed up as researchers on that specific project. Final outputs can be released beyond the project team. No counts of less than ten are permitted either for intermediate or final outputs and in this the NILS differs from the ONS LS. This restriction is policed by the NILS-RSU which also checks for factual errors in the ways that the NILS has been cited or described. At least one output has to be longitudinal and at least one output also has to have relevance to health and social care. Researchers are also strongly encouraged to pursue policy relevance and dissemination. The NILS-RSU keeps a record of intermediate and final outputs (or the links to access them) are made available on the NILS-RSU website. Projects are not kept live indefinitely. From the start, they have a fixed end date and this can be extended up to three times at the discretion of the RSU but if the researcher wants to extend it then further, approval must be sought from RAG.

9.5 Toward UK level LS data sets

An obvious question from an outside perspective is why there is not a UK level LS data set. The simple answer is that whilst the three studies are conducted within parts of the UK, they remain legally separate and cannot be easily commingled. A practical constraint lies in the fact that all three studies are accessible only via secure arrangements: in order to create a common data file (ignoring differences in sample size and content) it would be necessary for at least two of the studies to permit export of their data to the third study (or for all three to be centralized at a fourth location). This would not be consistent with the secure storage and access conditions of the data.

There are, however, a number of different levels of integration that can be considered: advances have made in a number of ways, and it is possible to speculate about the potential for further integrative work. For example, a relatively simple way of aiding understanding of population dynamics in each study would be to allow the partial tracing of members in administrative data in other parts of the UK. Whenever a new census is linked, it will be the case that there are a number of members who had been previously observed in one or more earlier censuses, but were not captured in the census being linked, and for whom there is no administrative record of death or embarkation. One plausible explanation is that the member has moved to another part of the UK. Thus, it would be useful for each agency (ONS, NRS, NISRA) to be able to transmit a set of minimal identifiers (such as an NHS number) to the other two agencies, who could then respond for each person identified indicating whether there is any administrative data to suggest that that person was resident in their respective territories at the time of the census.

We describe below three developments that are intended to both encourage wider use of the LSes and also to make cross-LS analysis easier. The CALLS data dictionary is a catalogue of variable information which pools metadata across all three studies; e-Datashield is an analysis technique that allows statistical operations to be carried out on multiple data sources without requiring those sources to be located in the same place.

Combined metadata: the CALLS data dictionary

The data producers for each of the three studies also maintain an ecosystem of supporting materials for their own study, and included amongst these are detailed data dictionaries, which list each field in the data tables, and give information about coding. The data dictionaries are invaluable resources for carrying out analysis. As part of the CALLSHub development plans, a combined data dictionary was developed which provided a single metadata repository providing information about all three studies. Whilst the separate dictionaries have all been developed to serve a similar purpose, they have different metadata structures and different implementations and therefore creating a single integrated dictionary is not simply a case of merging together the separate dictionaries.

The integrated dictionary therefore provides both a uniform way of querying the metadata, and also a consistently formatted set of results. More importantly it is designed to enable cross-study work, by allowing a user (or potential user) to determine whether a given variable exists in multiple studies. Whilst the query entry box encourages simple terms to be entered, it also supports wildcard characters and a number of Boolean modifiers, allowing advanced users to construct more complex queries. Advanced queries can include or exclude particular search terms, and can also allow the user to supply alternate search terms, should thematically similar variables be known to have different names in different studies.

Key to the development of the integrated dictionary has been the production of similarity scores for pairs of variables. For relevant variables, the search results will include a ‘Similar’ column, which give guidance as to whether related variables are similar or not.

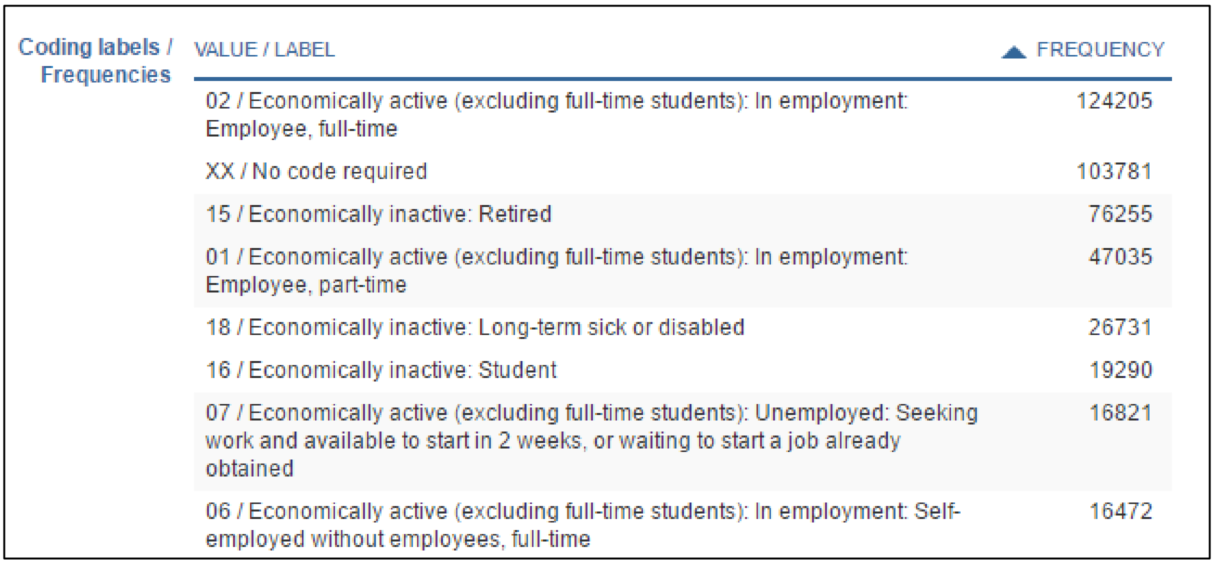
Error! Reference source not found. shows the set of possible scores that are reported. For each pair of variables (the variable currently being reported, and a potential equivalent) scores may vary from 1 (wording is similar, but question responses are not compatible) to 8 (question wording and responses are identical or near identical). It should be noted that the scoring is necessarily a broad brush approach – it is still contingent on the researcher to look closely at the variables involved (and to seek assistance from the research support teams if needed), but the intention with the similarity scoring is that it is possible to do an initial assessment of whether combined analysis is feasible or not.

Table 9.1 Similarity scores in the CALLS integrated data dictionary

Score	Meaning
0	No match found in other LSs, but some guidance notes given
1	Question wording similar, but categories incompatible
2	Question wording similar, categories may be aggregated to a common basis
3	Question wording similar, only minor differences in categories
4	Question wording similar, categories identical/near identical
5	Question wording identical, but categories incompatible
6	Question wording identical, categories may be aggregated to a common basis
7	Question wording identical, only minor differences in categories
8	Question wording identical, categories identical/near identical

Following on from the cross-study comparison, the similarity scoring can also be applied to cross-census consideration within the same study. Again, this allows an initial exploration to be done prior to potential research, to determine whether an apparently similar variable in different censuses (within the same country) can in fact be validly compared in analysis.

The similarity scores were developed in order to support users with one typical question (‘Are these variables the same?’) that might be asked when considering applying to use the data. A second common question is to ask whether there are sufficient numbers of people with a given characteristic to make analysis feasible, especially when those people are to be further disaggregated by other characteristics. A second development of the integrated dictionary has to capture and store frequency information for certain variables. An initial group of core variables has had such information added, with plans to extend the frequency data over time. **Figure** Error! No text of specified style in document.**.1** shows part of the output of the data dictionary for a sample variable ('ECOP1' – Economic activity in the NILS 2011 members table). The image shows the variable values (as included in data for analysis), the associated text labels, and finally an observed frequency in the sample data. Here, the labels have been re-ordered in frequency rank. Frequency observations are reported as being within a given range when small values might otherwise breach publication thresholds.



Coding labels / Frequencies

VALUE / LABEL	▲ FREQUENCY
02 / Economically active (excluding full-time students): In employment: Employee, full-time	124205
XX / No code required	103781
15 / Economically inactive: Retired	76255
01 / Economically active (excluding full-time students): In employment: Employee, part-time	47035
18 / Economically inactive: Long-term sick or disabled	26731
16 / Economically inactive: Student	19290
07 / Economically active (excluding full-time students): Unemployed: Seeking work and available to start in 2 weeks, or waiting to start a job already obtained	16821
06 / Economically active (excluding full-time students): In employment: Self-employed without employees, full-time	16472

Figure Error! No text of specified style in document.**.1** Partial screenshot of CALLS data dictionary

e-Datashield

Being able to compare the different parts of the UK or simply to increase the sample size available to a researcher, makes the combination of LSs studies an attractive option. Comparison of results between the different studies may be carried out by running separate analyses in the relevant safe haven and comparing the published reports (e.g. Popham and Boyle, 2011). This approach has several disadvantages. One can never be sure that the data sets and variables, which are nominally the same, are really comparable. An analysis that adjusts for covariates in each individual agency will not be identical to what one would obtain if the raw data were pooled. Tests for study-by-covariate interactions are not readily carried out from published reports. A similar situation has arisen in the analysis of genomic data, where a pooled analysis of small individual studies is required for adequate inference, but the individual centres do not wish to share their data.

The DataSHIELD system[1] was developed in response to this and implements a joint analysis by linking the computer in each centre to an analysis computer (AC). The AC holds no raw data, but receives summary statistics from each of the individual studies, combines them, and passes the combined summaries back to the individual centres. This allows joint analyses such as generalised linear models (GLMs) to be fitted by iterating this exchange of summary statistics. The interface between the AC and the other centres prevents any raw data being exchanged. Because security concerns would not allow LS centre computers to be linked in this way, an adapted procedure by exchanging summaries between agencies by email has been adopted by the LSs. Routines in R have been developed to allow such analyses to be carried out via the E- DataSHIELD protocol.

Synthetic data

A further recent innovation has been the development of two types of synthetic data. Firstly, a synthetic 'spine' dataset has been developed based on data and observations that are open licensed and thus easily disseminated. Secondly, a set of tools has been produced which can derive a synthetic set of output data from a sensitive (and non-shareable) set of input records. These approaches are designed to encourage new users and to make analysis more practical for existing users. In the longer term, it is plausible that these techniques could be used to produced data sets which could be held in the same location, thus enabling easier UK-level analysis.

[1] www.datashield.org

The synthetic LS ‘spine’ dataset (Dennett *et al.,* 2015) includes transitions of key demographic variables included in the national LSes. It was created using the 2011 England and Wales Teaching SAR dataset, available from the Office for National Statistics and a series of 2011 back to 2001 transitional probabilities taken from the England and Wales LS. A new LS-like dataset with plausible distributions was developed by firstly estimating the numbers of individuals in particular age groups undergoing each longitudinal state transition and then allocating transitions to the appropriate number of individuals in the SAR micro-dataset. Transitions applied include general health, marital status, religion, and approximate social grade. In addition, live births to females were estimated and added, and likelihoods of death over the ten-year period were modelled. The initial synthetic data set was produced for England and Wales, and was published together with the R source code for all algorithms applied. The same approach has been used to develop a similar data set for Scotland, with a Northern Ireland version also in preparation.

The ‘synthpop’ project was led by SLS-DSU (Nowok *et al.*, 2015) and produced an R package to generate bespoke synthetic datasets for individual research projects. The data are protected by removing sensitive variables and replacing them with synthetic versions. Replacements for categorical or continuous variables are generated by drawing from conditional distributions fitted to the original data using parametric or classification and regression trees models. Users can request synthetic versions of the data they request from the LSes for use outside of the secure microdata laboratories, subject to confirmation by the data holders. These synthetic versions will allow for simple tasks such as the refining of analysis scripts to be carried out more easily and we are confident that the synthetic data will be good enough to produce analysis results very close to those that would be carried out on the real data. After developing analyses on the synthetic data users will have the option of having them repeated on the actual LS data sets; indeed users are advised that they should not claim statistical validity in their results unless and until they have repeated their analysis on the true LS data.

9.6 Conclusion

The chapter has described three closely related sets of data: the ONS LS, the SLS and the NILS. These are all complex resources, and thus the differences between them are at times subtle but significant. Together, they form a very rich family of research data. Longitudinal data sets have also been developed in other countries in the world, and if the UK is to remain confident that its own studies are of a gold standard, then we must continue to maintain and improve them. With complex data it is fair to say that analysis will always require researchers to have a background of suitable and specific training. The aspects of the studies which can be improved lie in ease of access and usability, and these are areas where the UK community cannot rest on its laurels.

We still see no real feasibility of a UK LS in the short term, although we remain hopeful that a legal route to permit access to multiple data sources is one day found. In the absence of the easy ability to work on multiple LSes directly, a more hopeful avenue may be via synthetic data. If synthetic data derived from 'real' data are allowed out of safe settings (whether that be to the researcher's desktop or to a virtual secure environment) then it will become easier to carry out multi-country analysis.

When considering the future, we can also look towards future data sets. The next census will take place in 2021, after which the LSes will be extended with additional census data. This will give a total time span, especially for the ONS LS, that remains internationally impressive: six censuses spanning fifty years. It is interesting to observe that in 1971 the median population age (UK, rather than England and Wales) was 34.1 (Smith *et al.*, 2005). Moving these people forward 50 years gives an age of 84.1, which is greater than current (national) life expectancy (and quite a bit past life expectancy in 1971) – so, the older half of the 1971 sample have passed their life expectancy. We currently have no knowledge of the final set of questions that will be in the 2021 Census round, but can note that there are some variables that were new in 2011, and we will have our first chance in the 2021 sample to see whether and how these have changed (assuming that they are asked again). At the same time, we are hopeful that the emergence of the ADRN might facilitate wider linkage of administrative data with the LSes, further enriching the research potential they offer. Just as there are three separate LSes, with distinct characteristics, so there are four administrative data research centres (there are separate centres in England and in Wales) and thus progress is likely to be at different speeds in different contexts.

References

Acheson, D., (1998). Independent inquiry into inequalities in health report. The Stationery Office, London

Black D., Morris J., Smith C., and Townsend P. (1980) Inequalities in health: report of a research working group . Department of Health and Social Security, London

Blackwell, L., Lynch, K., Smith, J. and Goldblatt, P. (2003) *Longitudinal Study 1971–2001: Completeness of Census Linkage,*. Office for National Statistics, London.

Buck, N.. and McFall, S. (2011) Understanding Society: design overview. *Longitudinal and Life Course Studies, 3(1)*: 5-17.

Dennett, A., Norman, P., Shelton, N. and Stuchbury, R. (2015) A Synthetic Longitudinal Study for the United Kingdom, CASA WORKING PAPERS SERIES 201.

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

Office of Population Censuses and Surveys (1973) *Cohort studies: New Developments*. Studies on Medical and Population Subjects no. 25, HMSO, London.

Lynch, K., Lieb, S., Warren, J., Rogers, R. and Buxton, J. (2015), *Longitudinal Study 2001-2011: Completeness of Census Linkage, Series LS No. 7*, Office for National Statistics, Titchfield.

Marmot, M., Allen, J., Goldblatt, P., Boyce, T., McNeish, D., Grady, M. and Geddes, I., (2010). Fair society, healthy lives: Strategic review of health inequalities in England post-2010. The Marmot Review, London

Nowok, B., Raab, G.M. and Dibben, C. (2015). synthpop: Bespoke creation of synthetic data in R. Package vignette http://cran. r-project. org/web/packages/synthpop/vignettes/synthpop. pdf

Plewis, I. (2004) *Millennium Cohort Study First Survey: Technical Report on Sampling* (3rd. Edition). Centre for Longitudinal Studies, Institute of Education, University of London, London.

Smith, C., Tomassini, C., Smallwood, S. and Hawkins, M. (2005) The changing age structure of the UK population. In *Focus on people and migration* . Palgrave Macmillan, UK.

Turner, A., Drake, J., Hills, J. and Turner Commission, (2005). A new pension settlement for the twenty-first century. The second report of the pensions commission, The Stationery Office, London

Wadsworth, M., Kuh, D., Richards, M. and Hardy, R. (2006) Cohort profile: the 1946 national birth cohort (MRC National Survey of Health and Development). *International Journal of Epidemiology*, 35(1):49-54.