ESRC Centre for Population Change Report

Linking longitudinal studies of ageing with administrative data

First interim report

Summary of health & retirement studies' current data linkages

Dr Anne Gasteen, Dr Elaine Douglas, Professor David Bell

ESRC Centre for Population Change, University of Stirling

Contact: a.s.gasteen@stir.ac.uk

Acknowledgements

The work was funded through the ESRC Centre for Population Change (CPC) grant number ES/R009139/1.

ESRC Centre for Population Change

The ESRC Centre for Population Change (CPC) is a joint initiative between the Universities of Southampton, St Andrews and Stirling, in partnership with the Office for National Statistics (ONS) and the National Records of Scotland (NRS). The Centre is funded by the Economic and Social Research Council (ESRC) grant numbers RES-625-28-0001, ES/K007394/1 and ES/R009139/1.

The views and opinions expressed by authors do not necessarily reflect those of the CPC, ESRC, ONS or NRS.

The ESRC Centre for Population Change Report Series is edited by Teresa McGowan; t.mcgowan@southampton.ac.uk

Website | Email | Twitter | Facebook | Mendeley

Contents

Executive Summary

- The Linking Longitudinal Studies of Ageing with Administrative Data project is an ESRC Centre for Population Change (CPC) project. The project remit is to evaluate the administrative data linkage of the surveys belonging to the Health and Retirement Studies (HRS) family on the Gateway to Global Ageing (G2G) platform. It aims to provide a detailed understanding of linked studies' content, use, challenges and value added.

- The HRS' common aim is to better understand the issues affecting older people in their national populations, tracking and identifying changes over time. Cross-country comparisons are enabled by the G2G's production of harmonized datasets for the studies. This report provides an initial overview of current and planned HRS survey linkage based on publicly available information and documentation on individual study websites and the G2G platform.

- Linkage of routinely collected administrative data to longitudinal surveys can: provide better quality, more accurate data, reducing respondent burden and survey costs, identify bias and maintain representativeness. There are, however, substantial challenges to linkage viz: complex data ownership, ethical and legal issues resulting in lengthy application and approval processes, inflexible access protocols, considerable data cleaning, linkage methodology and accuracy.

- Ten of the eighteen HRS family surveys currently hosted on the G2G platform were found to be linked to administrative data. In general, the HRS family of surveys do not appear to be as advanced with respect to administrative data linkage as other longitudinal studies. The exception to this is the US HRS and, to a lesser extent, the ELSA and SHARE studies.

- Three initial observations may be made. Firstly, where linkages exist, these are often intermittent and can become dated by the time the linked data are available. Secondly, the development of linkage across the different studies is piecemeal and generally not uniform; surveys are linked to different types of data depending on data access conditions within different countries and study interests. Lastly, microlevel data linkage appears to be most successful and/or consistent where this has been built into the development and design of a survey at the outset and when there is political or legal weight behind this.

1. Introduction

Financial (income, taxation, benefits, pension), health (GP and hospital records) and education (school, college and university) data for individuals are routinely collected, nationally or locally, for administrative purposes. The increased use of digital records over the last twenty to thirty years has rendered them accessible for research purposes. On their own, individual administrative databases have been used to provide descriptive and/or trend analyses for field professionals and policymakers; for instance, in the analysis of the aetiology of diseases or in the setting of educational attainment targets. Their linkage to longitudinal studies can add particular value to lifecourse analyses, providing more accurate insights into the interaction of socio-economic characteristics and the outcome trajectories of individuals. For the worldwide collection of Health and Retirement Studies, arguably, such linkage is even more valuable. Such studies allow the inter-play between income, education and health to be traced out over the course of individual lifetimes into older age outcomes and potential needs, thereby providing intelligence as to the current and future socio-economic challenges that societies face against a background of (generally) ageing populations and a better-informed evidence base for policy interventions.

This is the first interim report for the Linking Longitudinal Studies of Ageing with Administrative Data project. Its remit is to evaluate the administrative data linkage of the surveys belonging to the Health and Retirement Studies (HRS) family on the Gateway to Global Ageing (G2G) platform. The aims of the project are to:

1. Comprehensively map the current, proposed and potential future linkage of the HRS surveys
2. Compare surveys' linked data content through a systematic review of the metadata
3. Undertake a review of the HRS' outputs (papers) using linked administrative data that are housed in the G2G publications' repository
4. Assess why and how surveys' linkage differs with respect to ethical consents and access conditions
5. Assess the *value added* of the administrative data linkages and their potential future role in the development of longitudinal studies of ageing

Aims (1) linkage mapping and (2) comparison of linked content, will be supported by a survey of study Principal Investigators (PIs); this is necessary and important not only to gauge future linkage plans but because current survey linkage documentation is often incomplete. Aim (3) the review of

HRS' linked administrative data outputs will be further underpinned by a survey of researchers using the linked studies to inform the evaluation of access permissions and procedures.

This report provides an *initial* overview of current and planned HRS survey linkage based on publicly available information and/or documentation on individual study websites and the G2G platform. The reported linkage information will be reviewed and built upon as the project proceeds, to provide a more detailed understanding of linked studies' content, use, challenges and value.

2. Background

There are now a substantial number of established HRS studies around the world. The G2G platform is a free public resource that aims to promote cross-national and longitudinal studies of ageing by facilitating discoverability of, and access to, these studies, as well as producing harmonized datasets for comparative research. The core content of the HRS family surveys includes demographic, health, health service, employment and other socio-economic variables as shown in Table 1 below. Physiological biomarkers, including height, weight, waist circumference, blood pressure, pulse rate, and peak expiratory flow are collected by most of the HRS family of surveys (the exceptions being the Korean, Japanese, Thai, Scottish and Malaysian studies, see Appendix Table A1). Questions about public and private health insurance, out-of-pocket medical expenses, health care utilization (hospitalization and visits to general practitioners), preventative care, medication, dental care, and long-term care are asked by all surveys, enabling policy effects on health care utilization to be evaluated (Lee, 2010).

Demographic	Education, marital status, age, resident, birth year, birth month
Health	Cognition, disease, depression, injury, physical functioning, physical measures, health behaviours
Health Services	Insurance, utilization, expenditure, out-of-pocket spending
Work & Employment	Employment status/history, labour force, earnings, disability, retirement, type of work, pension
Economic Status	Earnings, asset income, government transfers, pension, financial assets, housing, non-financial assets
Family Structure & Social Network	Parents' information, household structure, family exchange, family support, social participation

Table 1: Health & Retirement Studies: Core Content Areas

Source: Gateway to Global Aging [1]

[1] Gateway to Global Aging Data, Produced by the Program on Global Aging, Health & Policy, University of Southern California with funding from the National Institute on Aging (R01 AG030153)

In an earlier review that covers many of these longitudinal datasets, Kaiser (2013) identifies the main benefits, such as: providing a much clearer picture of the issues affecting ageing societies in different countries, generating significant research and new lines of inquiry, establishing a baseline from which governments and researchers can project and monitor age-related issues in their national populations and make cross-country comparisons.

The oldest of the studies is the US Health and Retirement Study (HRS), funded by RAND and the US National Institute on Aging. Subsequent studies are broadly modelled on the HRS for harmonization purposes. Many of these receive funding from the NIA and, therefore, comparability with the HRS is perceived as a benefit during the application process. The Gateway to Global Ageing creates harmonised HRS family datasets modelled on the *RAND HRS*; a user-friendly version of a subset of the US HRS created by the RAND Center for the Study of Aging to increase researcher accessibility. The HRS contains cleaned and rendered, standardised and derived variables that have consistent naming conventions (Beaumaster et al, 2018). The harmonized datasets are produced by re-defining variables, aligning these as much as possible to the RAND HRS and adopting its variable naming conventions. All study waves are combined to give one record per individual. A country-specific variable is included and there are spouse versions for most variables. At present, administrative data linkage would appear to occur only at source for the original HRS family surveys; the harmonized datasets available on the G2G platform do not appear to be linked to country-of-origin administrative data. Table 2 shows the HRS Family surveys for which the G2G has produced harmonized datasets and *sister* studies for which information is hosted on the platform but, as yet, there are no harmonized datasets.

HRS Family Surveys with Harmonised Datasets	
Survey	**Start year**
Health & Retirement Survey (HRS)	1992/93
Mexican Health and Aging Study (MHAS)	2000/01
English Longitudinal Study of Ageing (ELSA)	2002/03
Survey of Health, Ageing & Retirement in Europe (SHARE)	2004/05
Costa Rican Longevity & Healthy Aging Study (CRELES)	2004/05
Korean Longitudinal Study of Aging (KLoSA)	2006/07
Japanese Study on Aging and Retirement (JSTAR)	2006/07
The Irish Longitudinal Study on Ageing (TILDA)	2010/11
China Health and Retirement Longitudinal Study (CHARLS)	2010/11
Longitudinal Aging Study in India (LASI)	2016/17
HRS Family Sister Surveys (no Harmonised Datasets)	
Survey	**Start year**
Indonesia Family Life Survey (IFLS)	2007/08
Study on Global Ageing and Adult Health (SAGE)	2010
Health, Aging, and Retirement in Thailand (HART)	2011
Northern Ireland Cohort for the Longitudinal Study of Ageing (NICOLA)	2014
The Health and Aging Study in Africa (HAALSI)	2015
The Brazilian Longitudinal Study of Aging (ELSI)	2015
Healthy Ageing in Scotland (HAGIS)	2016 (Pilot)
Malaysia Ageing and Retirement Survey (MARS)	2018/19

Table 2: HRS Family Surveys on the Gateway to Global Aging Platform

3. Data Linkage

3.1 Common Challenges of Longitudinal Surveys

The HRS family of surveys have a common aim: "to better understand the multifaceted lives of older individuals and their families and to track and identify changes over time." (Lee, 2010, p3). The studies face the same challenges as all longitudinal surveys. In addition, those studies that want to augment their survey data with administrative data face further challenges.

The Kaiser 2013 review of these studies identified four main challenges that, generally, they share with other longitudinal studies. Firstly, funding was consistently mentioned as an issue by studies, impacting on both survey intervals and their content (see, for example, Weir & Ofstedal, 2016). Secondly, ethnic minority coverage and sample size has been problematic for some studies (for instance ELSA). Thirdly, the level of detail on specific health outcomes may not be as great as for tailored studies. For instance, biomarkers and physical health measures may not be collected as regularly as would be desirable as they are costly and burdensome to collect. Fourthly, international comparability needs to be factored into the development phase of studies.

Other challenges that the studies face are: the recruitment and retention of respondents, the expansion of cohorts to widen the age range (which, while desirable for more robust tracking of phenomena, is more costly), cultural sensitivities and regional disparities for each country (Kaiser, 2013), adapting questionnaire content to changes in society (e.g. digitisation or changed patterns of ageing) and increasing investment in technological advancement to make use of the new technological possibilities for surveying individuals (ESRC, 2017). In the ESRC's strategic review of longitudinal studies, it is argued that the most pressing need is to increase the analysis potential of such surveys and their questionnaire programmes by enriching them through linkage to administrative data sources, and by increasing harmonisation between surveys internationally (*ibid*).

3.2 Harmonization

The HRS family of surveys have been developed in line with the US HRS to share a common research design and collect core, comparable micro-data (Lee, 2010). This facilitates the production of harmonized datasets from the studies for international comparison purposes. It is generally recognised that the HRS family of surveys are much more advanced in terms of data harmonization than other longitudinal studies. This is a result of the work carried out by the G2G to produce accessible, harmonized datasets. The scientific value of these harmonized HRS family datasets will grow as the longitudinal data accumulate enabling the longitudinal analysis of changes in health and labour force behaviour dynamics (Lee, 2010).

3.3 Linkage Benefits

The main benefits and limitations of using administrative data in their own right for research purposes are summarized in Table 3, together with the opportunities and challenges that linking administrative data to survey data offers. The linkage of longitudinal surveys with different, routinely collected, individual or aggregate, administrative data has enormous potential to enhance and expand survey datasets. Linkages offers the potential to add information that cannot be obtained or cannot be easily or accurately collected from survey respondents. Data collection is non-intrusive, more efficient and no additional collection costs are incurred. The data are regularly updated with total coverage of the specific administrative population, providing more accurate, timely and comprehensive information. By providing further information, linkage can be used to increase the completeness of survey data and correct respondent bias, resulting from missing or inaccurate data on, for instance,

Administrative Data	
Advantages	**Limitations**
No additional costs for collection and non-intrusive	Costs of making data research-ready
Regularly updated	Lack of metadata
Total coverage of administrative population	Legislative barriers
Potential controls and booster samples post hoc	Data controllers as data gatekeepers
Reliable at small area level enabling better place-focused policy making	Coverage of population subject to policy changes
Timely data capture enables time series to explore trends	Content limited to that collected specifically for administrative purposes
Linked Longitudinal Survey & Administrative Data	
Opportunities	**Challenges**
Improved representativeness across time, geographies & sub-populations	User training
More cost effective	Resources
More accurate, timely and comprehensive information	Access
Increased reliability and validity	Consent
Enables better theory and policy driven research	Legislation
Can test for range of factors to understand complex social phenomena (e.g. mobility)	Data Management

Table 3: Advantages and Limitations of Linking Longitudinal Survey and Administrative Data

Source: (Adapted from: ESRC, 2017, Figure 3, p18)

hard to recall events. If the linkage can enable respondent information to be replaced with better quality, external source data, this will increase the reliability and validity of survey analyses. The use of more accurate, reliable, timely and comprehensive data enables theory and policy driven research. Linkage may be more cost effective, allowing routine information to be collected from the administrative data, reducing the time needed to conduct surveys and easing the burden on

respondents. In turn, this can create space for surveys to concentrate on the collection of objective-specific information that can only be gathered in this way.

Linkage and/or the use of administrative data as a population spine for longitudinal surveys, offers the potential to deliver policy relevant insights more quickly while ensuring representativeness. Longitudinal surveys' representativeness across time, smaller geographies and sub-populations may be improved through administrative data linkage. As the data effectively provide census type coverage, analysis of them is reliable at small area levels enabling better place-focused policy making. Their use makes it easier to devise potential control groups and post hoc booster samples. The representativeness of survey data may be increased by enabling sample subpopulations to be nested within a wider administrative population. This makes it possible to assess the extent of cross-sectional non-response bias to be assessed and devise sample population weights, facilitating the population level analyses of subpopulations (ESRC, 2017).

Moreover, survey sample creation using an administrative population spine can ensure representativeness. Linkage provides a means of increasing established surveys' coverage of subpopulations that may be underrepresented in traditional data collection processes but covered in an administrative database. Switching to an administrative population spine to refresh survey samples can facilitate their representativeness. Using such a spine would make cohort comparisons, sample boosts and the development of new cohorts to address emerging research issues easier. In turn, it offers the potential to facilitate linkage to other administrative data sources by providing more streamlined access.

Combining survey and administrative data can enable the analysis of complex social phenomena that are determined by a range of factors. For instance, it may be possible to establish causal inference from medical records and model health outcomes and the data are likely to be of higher quality than that obtained from surveys with self-reporting. Timely data capture enables the exploration of trends using time series' techniques.

3.4 Linkage Challenges

As shown in the bottom panel of Table 3, in addition to the challenges that all longitudinal surveys face, studies that want to augment their survey data with administrative data face further challenges. Administrative data are collected for administrative purposes, not to facilitate research, and data ownership can be complex.

A major drawback of using administrative data for research purposes is that they are not collected for this; they will tend to comprise the minimum necessary data to establish the administrative function and population coverage may be impacted by changes in policy. There are no research-sympathetic metadata and administrative data entry can be highly variable. Data may be entered inconsistently by different individuals across years. More superfluous data from an administrative perspective, whilst collected, may not be entered at all. Duplication may be present in a proliferation of administrative codes. The result is that much time can be spent trying to both rationalise and complete an administrative dataset before derived variables of interest can be created. Often, inconsistencies in administrative datasets may not be detected until midway through a rendering process (Gasteen, 2019). Where administrative data are irretrievably missing or incomplete, appending them to a longitudinal survey may reduce the sample size and representativeness for potential analyses.

Data ownership can be complex. For instance, in the UK, there are a wide range of data owners with some datasets covering the whole of the UK, while others are devolved and some are regional (ESRC, 2017). Wide ranging ownership of administrative datasets in different countries can make longitudinal studies' negotiation of access challenging in terms of the time and resources needed. Moreover, ownership and control may not be the same thing, necessitating negotiation for access with an organisation that is not the data owner. Usually, government departments own the administrative datasets of primary interest to longitudinal studies. However, there may be further complexities of ownership of records within government departments, reflecting operational divisions. Changes in ownership and/or policy can affect access to administrative data. For instance, changes to legal frameworks may reduce the size of datasets, causing some information that was previously shared to be removed or withheld going forward (Gasteen, 2019). Changes in policy may also affect the population coverage of an administrative dataset as, for instance, the categorisation or eligibility of individuals changes.

Application, approval and access processes can be very time consuming and often at odds with project and funding timelines (Brett & Deary, 2014, Dattani et al, 2013). The ease of navigation through each of these stages will be determined variously by cultural, ethical and legal framework differences between countries and different interpretation/implementation *within* countries. For instance, the interpretation of common data protection legislation and the ensuing General Data Protection Regulation (GDPR) requirements varies between countries in the EU. In the UK, different

government departments may require different undertakings from studies wishing to access their data, resulting from differing interpretations and applications of GDPR. Privacy and security issues dominate application, approval and access processes, producing a fundamental tension between confidentiality and analytic convenience and scope.

The issue of consent arises on two fronts, with both dataset owners and, potentially, survey respondents. Firstly, the consent of the data owner is required through a data share agreement. In turn, the data owner may or may not require that the study has the consent of respondents, depending on the type of data and interpretation of data protection legislation. The consent of respondents may or may not be necessary, again depending on the type of data and interpretation of data protection legislation. There has been growing concern that linkage may increase the risk of identification, in addition to the risk from the simple release of administrative data (ESRC, 2017). Moreover, a grey area appears to be emerging as to whether individual consent for linkage to contextual administrative data is required (e.g. for adding local authority areas or geographical deprivations indices to survey records).

While GDPR effectively requires that respondent consent should be explicit and documented, the legal necessity for such maybe misunderstood as it is not always the appropriate lawful basis by which to meet legislative requirements (ESRC, 2017, Annex K). For instance, universities may rely on public task as the legal basis for processing personal data for research purposes.[2]

Respondent consent depends on the degree of trust that individuals have in the organisation/s responsible for the data (ESRC, 2017). Consent by design, where consent is built in at a survey's recruitment stage and is a condition of participation, would appear to offer a stable platform for conducting linkage updates overtime. Consent that is future proof may be considered good practice, enabling linked data to be used in ways that may not have been conceived of at the time survey conception (OECD, 2016). Consent by design is not necessarily future proof, however, as expectations, legal interpretation and/or policy may change over time, requiring consent renewal. Having to undertake periodic renewal of consent may be resource intensive for studies and confusing for respondents. To overcome this, some studies have sought to use legal consent alternatives or data processing procedures to produce de-identified respondent information (ESRC, 2017, Annex K).

[2] Lawful basis for processing | ICO, accessed 31/01/21.

Where respondent consent is required, this has the potential to impact on survey response rate and representativeness, reducing statistical power and introducing bias (Kho et al., 2009, Knies et al. 2012). Consent rates are likely to vary by socio-economic group and, possibly, also by health status. Context can be important too as consent may vary according to data source, with individuals consenting to linkage to one type of administrative data while withholding consent for another. Unless consent by design is all encompassing of proposed linkages, this may result in incomplete response across the survey sample.

Once data sharing agreements are in place, there may be direct and indirect costs of access. Data controllers may charge for access to their data infrastructure. If access is provided through dedicated, *safe haven* centres that researchers have to travel to, rather than through secure, remote desktop access, then indirect time costs are incurred.

Linkage methodologies and processes are also important as linked data may be subject to potential biases arising from, variously: consent, missed links, incorrect links (ESRC, 2017, Harron et al., 2016, Sakshaug et al., 2017). It has been shown that linkage errors can vary with demographic characteristics: age, sex, place of birth, ethnicity, socio-economic status, and event dates (Bohensky et al, 2010). Appropriate analysis of linked data depends on knowledge of how accurate the linkage is and requires relevant *paradata* for end-users, describing how the data are obtained, the potential distribution of linkage errors and its randomness. [3] Paradata, however, are often not available or of inconsistent quality (ESRC, 2017, Wagner et al, 2013), making the assessment of linkage quality problematic, necessitating the use of indirect evaluation methods.[4]

The ideal linkage key is a unique identifier, for instance, a UK National Insurance (NI) number. This allows *deterministic* matching of individual records to take place, that is where there is exact agreement on the unique identifier across datasets. Linkage based on deterministic matching, using a unique identifier, should mean that the potential for bias arises only from consent and not the matching process. Deterministic matching can also be conducted on specified sets of partial identifiers (eg surname, sex, postcode). The process can be modified to allow for small differences in partial identifiers by using a sequential set of rules (algorithm). Deterministic matching is designed to avoid

[3] If linkage errors are not random, linked datasets may suffer from bias and no longer be representative of the population under investigation.

[4] Possible methods might involve: the comparison of distributions of variables from a linked dataset with those from the original data source, or with a gold standard dataset created on the basis of obtained true linkage error, or estimates with actual values from an alternative, trusted data source, or the clerical audit of a linkage sample (the latter would necessitate access to the original data).

false matches but can be prone to missed matches, thereby introducing bias from missed links (Harron et al, 2017).

Probabilistic matching enables linkage in the presence of recording errors and/or without using a unique identifier. Weights are applied to a chosen set of identifiers to reflect their contribution to an overall *match weight*. The weights are designed to reflect the relative discriminative value of identifiers. For instance, agreement on date of birth provides more powerful evidence of a match than agreement on sex.

Computed match weights are compared to a threshold or cut-off value to classify records as links. The choice of threshold values is important; adjusting threshold values will alter the balance between the of number of false and missed matches, and, therein, the potential bias arising from incorrect and missed links. False and missed matches move in opposite directions as a threshold is adjusted. Increased threshold values are likely to result in fewer false matches but more missed matches. Decreased threshold values are likely to reduce the number of missed matches but increase the number of false matches (Harron et al., 2016).

The majority of linkage errors arise from discrepancies in identifiers between the databases to be linked. Incorrect matches (false positive errors) tend to occur more frequently when there are: a limited number of identifiers and/or their completeness is low, a high proportion of homonyms and large volumes of data (Coeli, 2015). Missed matches (false negatives) tend to be caused by incomplete or data entry errors: incomplete or incorrect information and typographical errors (*ibid*).

Additionally, econometric issues may arise in the analysis of individual survey records that are linked to contextual administrative data, where averages are either applied or substituted for individual values. Whilst aggregate values are often used in social and medical research, applying a group average to an individual can be misleading. For instance, one study has shown the variance in individuals to be up to four times greater than the variance in groups (Fischer et al, 2018).[5] The use of group averages can lead to fundamental model misspecification that cannot be dealt with simply by applying robust standard errors (King & Roberts, 2015).

[5] Problems in the use of aggregates may arise as a result of outliers distorting average values, the averaging of different populations or averaging across a skewed population.

The issue of linkage accuracy and ensuing problems of bias may by compounded by a lack of high-level, methodological training in the statistical issues that can arise as a result of linkage and a lack of appropriate support for such analyses in statistical software. The ESRC's 2017 Strategic Review posited that the creation of linked datasets was running far ahead of the capacity to analyse the linked data while accounting for these biases in a statistically rigorous fashion.[6]

4. HRS Family Surveys' Data Linkage

A review of study websites indicated that seven of the ten *harmonized* HRS Family surveys and three of the eight sister surveys have administrative data links as shown in Table 3. No explicit administrative data links were found on the remaining studies' websites. HRS, the longest established survey, is the most comprehensively linked study. It is linked to eight different sources of administrative data, followed by ELSA with six linkages and SHARE with four. HAGIS, one of the sister surveys, has been linked to two different sources of administrative data and currently shows the greatest potential for further linkage with six additional sources indicated. HRS family surveys' administrative data linkage is perceived to be less advanced than other longitudinal studies and, in the Kaiser review, was described by one principal investigator "plagued by problems of privacy/access" (Kaiser, 2013). Comparing the HRS Family linkages shown in Table 3 with those of UK longitudinal studies (Appendix Table A3), it is clear that, collectively, the former are not as comprehensively linked.

Individual survey linkage is discussed in more detail below while a summary of the key features of the studies is given in Appendix Table A1 together with a list of collaborating institutions and study websites in Table A2.

[6] ESRC (2017), *Longitudinal Studies Strategic Review*

Domain	Dataset	Health & Retirement Survey	Mexican Health and Aging Study	English Longitudinal Study of Ageing	Survey of Health, Ageing & Retirement in Europe	Costa Rican Longevity & Healthy Aging Study	Japanese Study on Aging & Retirement	The Irish Longitudinal Study on Ageing	Healthy Ageing in Scotland	The Health and Aging Study in Africa	Northern Ireland Cohort for the Longitudinal Study of
Population Registry	Civic Records / Census Data										
Health	Secondary & Community Care										
	Primary Care										
	Social Care										
Education	School										
	FE / HE										
Employment	Employment										
	Earnings / Tax										
	Benefits										
	Pension										
Criminality	Convictions / Cautions										
	Violent Death Count										
Spatial	Geo-spatial										
	Environment Exposures										
	Housing										
	Neighbourhood Services, Green Space										
Linkage Key		**Linked**		**Linked: Data type unclear**			**In progress / potential to link**				

Table 3: Data Linkage for HRS Family Surveys

4.1 HRS Linkage:

HRS, the oldest of the HRS Family surveys, was established by the US Congress 1990 Act, which directed the National Institute on Aging to create a survey-based study to support research and inform policymakers at the national level about health and retirement issues affecting the increasing population of older Americans. [7] HRS is linked at the individual level to economic, health and contextual administrative data viz:

- Social Security Administration (SSA)
    - Earnings
    - Benefits
- Centers for Medicare & Medicaid Services (CMS)
    - Medicare and Medicaid claims
- Veteran Affairs (VA)
- Employer-provided pension plans
- National Death Index
- Census data, community surveys, air quality, crime, food access, medical resource distribution, long-term care use and facilities

Waves have been comprehensively (i.e. sequentially) linked to SSA and (various) CMS data from the survey start in 1992 until 2018.

Access permissions: Access to HRS data including geographical variables and administrative records from Social Security and Medicare is restricted and requires special agreements (G2G, 2015). Otherwise publicly available HRS data can be obtained directly from HRS website.

4.2 MHAS Linkage:

MHAS aims to examine health and economic outcomes for a large representative sample of older Mexicans, comparing these with older Mexican-born immigrants/second generation Mexican-Americans in the US and Mexican migrants returning from the US. The study focuses on the ageing process, associated diseases and disability burdens (Kaiser, 2013) and the role of intergenerational

[7] https://hrs.isr.umich.edu/

transfers. MHAS was designed to have similar protocols to the HRS, enabling the comparison with Mexican-Americans based in the US. MHAS is linked to the following administrative data:

- National Institute of Statistics and Geography
    - Community level characteristics
- Mexican Ministry of Health
    - Community level characteristics
    - Municipal level mortality
- Mexican Ministry of Health / National Population Council
    - State level Seguro Popular enrolment (public health insurance)

MHAS linkages to date appear to be to contextual data only (background geographic, socio-economic or environmental data that are not specific to the survey respondent) and are not always consistent across wave dates but this may reflect a time lag in retrospective linkage across later waves. Wave One (2000/01) is linked to community-level characteristics from Population and Health Services Censuses. Waves One, Two and Three are linked to death count by cause data and Violent and Non-Violent Death Counts (www.mhasweb.org, G2G, 2015). There is no obvious individual consent.

Access permissions / conditions of use: Access to the linked contextual data is restricted. Otherwise, publicly available data availability and documentation are available in Spanish and English.

4.3 ELSA Linkage:

ELSA was designed to provide information on the dynamics of health, social wellbeing and economic circumstances in the 50-plus English population to inform policy making (www.elsa-project.ac.uk/about-elsa). It is modelled on the HRS. Linkage to administrative data was an initial objective of the ELSA study (Kaiser, 2013). The study website reports that survey 'respondents give permission to link their ELSA data' to administrative data from:

- Office for National Statistics
    - Death records
- National Health Service (NHS England)
    - Hospital Episodes Statistics (HES)
    - Mortality and cancer registration records (NHS)
    - Primary care records
- Department for Work and Pensions' (DWP)

    - Economic records including National Insurance contributions
    - Benefits
- Her Majesty's Customs and Revenue (HMRC)
    - Tax credit records

The Study Description (G2G, 2015) states in a discussion of the then, forthcoming Wave 7 that ELSA was to be linked to 'routinely collected primary [health] care records'.

Access permissions: Access to ELSA survey data collections containing sensitive data (viz: linked administrative records, disclosive geographical variables, data on genetic material) is restricted (G2G, 2015). For such datasets that are visible on the UK Data Service site, generally, special license access/secure access (the latter through the UKDS Secure Access gateway) is required. These datasets variously contain: Local Authority boundaries, Census Rural/Urban indicators, Population density Post Code sectors, Index of multiple deprivation indicators, lower/middle layer Super Output Areas, State pension age data, primary care data, sexual self-completion data.[8] Access to non-sensitive, publicly available ELSA data, including the Harmonised ELSA dataset produced by the G2G, can be obtained through the UK Data Service website.

4.4 SHARE Linkage:

SHARE was created in response to a European Commission call to examine the possibility of creating a European longitudinal survey of ageing and has legal status as the first European Research Infrastructure Consortium - ERIC (Kaiser, 2013). SHARE is a multidisciplinary, cross-national panel interview survey of the health, socioeconomic status, and social and family networks of individuals aged 50-plus (Lee, 2010). It aims to analyse how the living conditions of the over 50s are shaped by the interplay between health, economic and social factors in 27 European countries and Israel (*ibid,* G2G[9]).

SHARE administrative data linkage takes place through country specific projects. To date, three SHARE countries have carried out administrative data linkage projects: Germany, Denmark, and the Netherlands,[10] while plans are reported for projects in further six countries: Austria, Belgium,

[8] The data we collect | ELSA (elsa-project.ac.uk), accessed 03/01/21.
[9] https://g2aging.org/?section=surveyOverview#tab-content-0/, accessed 27/05/20.
[10] http://www.share-project.org/special-data-sets.html accessed 13/06/20.

Estonia, Finland, Luxembourg and Spain (Borsch-Supan et al., 2013). Record linkage to SHARE data was undertaken first in Germany, with the establishment of the SHARE-RV project, followed by the REGLINK-SHAREDK project for linking Danish SHARE data and, most recently, the Dutch record linkage project.

4.4.1 SHARE-RV

The SHARE-RV project began with a pilot in the third wave of SHARE and has been integrated as a standard module of the German SHARE questionnaire since (Waves 4, 5, 6 and 7). SHARE-RV links German SHARE data with administrative data held by German Pension Fund (DRV) conditional on respondents' consent.[11] The administrative data are comprised of two databases that are updated annually:

- A longitudinal panel database, beginning at age 14, that includes socio-demographic characteristics (age, sex, number and age of children, education), detailed work history information and all activities that generate public pension entitlements provided on a monthly basis
- A cross sectional database for retirees only that includes information on the calculation of pension benefits (Korbmacher et al, 2013).

4.4.2 REGLINK-SHAREDK

REGLINK-SHAREDK was completed in February 2017 by establishing the project as a national research infrastructure. The project has enabled Danish SHARE data, for Waves 1 to 6, to be linked to official health, labour market and demographic register data from Statistics Denmark and the Danish Health Authority. There is to be subsequent updating as new SHARE waves are released. Participating individuals were provided with information about the project in accordance with the Danish Act on Processing of Personal Data. A central personal identification number provides the mechanism for linking Danish SHARE data with individuals' administrative health records.

[11] http://www.share-project.org/special-data-sets/record-linkage-project/share-rv.html

4.4.3 SHARE Netherlands

In the latter half of 2019, the Dutch record linkage project completed the linkage of Waves 1 to 6 of SHARE Netherlands (SHARE NL) survey data to the administrative records of the Dutch national statistical office, Statistics Netherlands (CBS). SHARE NL datasets for Waves 1 to 6 can be linked at the individual level to data in the CBS Microdata Catalogue for respondents who consented to this linkage in Wave 5. The Catalogue contains the population registry and data on health, work, social security, income, and wealth.

Access permissions / conditions of use: General SHARE data (regular panel waves) are distributed by SHARE-ERIC to registered users to download free of charge through the SHARE Research Data Center. Access services are provided through two public data archives in partnership with the Central SHARE Coordination Team: the CentERdata Archive, Tilburg University, Netherlands, the Data Archive for the Social Sciences (a public data archive) run by GESIS Leibniz-Institute for Social Sciences in Cologne.[12]

German Administrative Data is provided by the Data Research Center of the German Pension Fund (FDZ-RV), via its website, as a separate dataset that can be linked to the German SHARE data. [13] To get access to the administrative dataset, researchers must register with the Center.

Danish SHARE linked data can be accessed remotely on a secure server at Statistics Denmark by researchers at Danish institutions. Researchers must be registered as users of SHARE data to gain access. Access details can be found on the website of the Danish SHARE-DK research consortium.
Dutch Administrative Data are analysed within a secured working environment at CBS Microdata Services. Researchers at institutions authorized by CBS can access this environment remotely, from any secure workplace via a secure Internet connection. Only statistical results may be exported. There is a charge for CBS' services. Institutions that are not authorized by can apply for Statistics Netherlands' approval. Detailed access information is provided at www.cbs.nl/microdata.
There is a two-stage process to apply for access to the linked SHARE data at Statistics Netherlands; researchers need to:

1. Agree to the SHARE conditions of use and signing the SHARE user statement.
2. Contact the Dutch Country Team Leader to apply for access to the linked data.

[12] http://www.share-project.org/data-access.html accessed 29/05/20.
[13] *Ibid*.

4.5 CRELES Linkage:

CRELES' main study objective is to examine the length and quality of life, and contributing factors, for elderly Costa Ricans who have an unusually high life expectancy for a middle-income country; life expectancy in Costa Rica is higher than in the United States. [14] CRELES has been linked to the following administrative data since Wave 4, using respondents' national ID cards:[15]

- Costa Rican National Death Index
- Birth records
- Voting records

Access permissions: Public access CRELES data are available for academic research use only via submission of an online request for access to the University of California, Berkeley.[16] It is possible also to request access to the linked administrative data.

4.6 JSTAR Linkage:

JSTAR is conducted for academic research and policymaking purposes, primarily to inform the development of a sustainable social security system in Japan through the collection of data on the economic, social, and health conditions of individuals aged 50 or over. This is seen to be more urgent in Japan as its population is ageing ahead of other developed countries and has the longest life expectancy at birth of any country.[17] The ability to link JSTAR to individuals' medical and long-term care use was built into the survey at the outset with the support and permission of the five sample local government municipalities (Adachi, Kanazawa, Sendai, Shirakawa, Takikawa) viz:

- The National Health Insurance (NHI) programme
    - A national mandatory plan that provides hospital and medical services for the self-employed and retired [18]
    - Linkage to this enables medical record access
- The long-term care insurance (LTCI) programme
    - Established in 2000 and covering all Japanese residents, to provide non-medical elderly care for those aged 65 plus

[14] http://www.creles.berkeley.edu/pdf/Methods_w3.pdf
[15] https://g2aging.org/?section=overviews&study=creles-2, accessed 24/11/2020.
[16] http://creles.berkeley.edu:1313/CRdata.pl
[17] https://en.wikipedia.org/wiki/Health_care_system_in_Japan, accessed 10/04/2020.
[18] Employers fund medical insurance for employees (*ibid*).

The programmes are run at the municipality level where the official records of medical and elderly care services' use are held. At the Wave 1, JSTAR respondents participating in these programmes were asked for permission to connect their survey data with the official, monthly service use records for these two programmes for the previous 24 months.

Access permissions: Access to data is for academic and statistical research use only. An application must be made to the Research Institute of Economy, Trade and Industry (RIETI) as the host institution. There are two levels of access: High (H) and Very High (VH) approval status. H status allows access to the anonymized, randomly selected data. VH status is required to access full sample datasets containing municipality-level geographic information (addresses of individuals at municipality level). VH level users can access these datasets at their own workstations but only through the Secured Access System for Elaborated Study (SACSES). This is RIETI's remote control system, designed to prevent leakage of information that could potentially result in the identification of individuals. [19]

4.7 TILDA Linkage:

The aim of TILDA is to understand the health, economic and social circumstances of the older Irish population and how the ageing process is influenced by the interactions of these factors.[20] The survey has the greatest breadth of physical, mental health and cognitive measures collected amongst longitudinal studies internationally (Kaiser, 2013). TILDA now has administrative data linkages from Wave 1 to the following:

- Health Service Executive
    - Prescription data
- General Registry Office
    - Death records

Access permissions / conditions of use: Publicly available, anonymised TILDA datasets can be accessed for research purposes through an on-site Hot Desk Facility. The data are available from either the Irish Social Science Data Archive (ISSDA) or the Interuniversity Consortium for Political and Social Research (ICPSR). A request form must be completed.

[19] https://www.rieti.go.jp/en/projects/jstar/index.html/, accessed 14/04/2020.
[20] https://tilda.tcd.ie/data/documentation/doc/TILDA%20Release%20Guide%20v4.1.pdf, accessed 30/1/2020

4.8 HAGIS Linkage:

HAGIS aims to improve the lives of older Scots by increasing the understanding of their lives to inform policy to meet their needs. Data were collected on the socio-economic circumstances and health of a sample aged 50 and over. The study was informed by the HRS family of studies to enable harmonised data and international comparisons. Administrative data linkage has been used to develop an innovative sampling frame and link survey data to health, social care, and education data.[21]

As part of the main interview, respondents were asked for their consent to link their data to various administrative datasets: NHS Health records, Social Care Survey, Education records, Department for Work and Pensions, Her Majesty's Revenue and Customs data.[22] Respondents were able to consent to some, none or all the datasets; consent rates for the specific datasets were:

- NHS Health records- 85%
- Dental records – 82%
- Social Care Survey- 81%
- Education records- 80%
- DWP and HMRC- 77%

The linkage process comprised three key stages (Douglas et al. 2015):

1. Respondent consent to link their survey data to administrative data
2. Approval from the Administrative Data Research Network (ADRN)
3. Approval from the administrative data controller

An application was made to the ADRN's centre in Scotland, the Administrative Data Research Centre (ADRC) Scotland, to access de-identified administrative data in its secure environment; this was approved in January 2016. The application was made on the grounds of feasibility, academic merit, public benefit and privacy impact.

The third stage involved an application for approval to the Public Benefits and Privacy Panel (PBPP) which undertakes information governance security of requests for access to NHS Scotland Health and Social Care data in Scotland. Permission was sought to i) screen the sampled households for eligibility using the administrative data held by National Records Scotland (NRS), and ii) link

[21] http://www.hagis.scot/research/the-development-of-hagis/, accessed 08/09/20.
[22] http://www.hagis.scot/data/, accessed 09/09/20.

respondents' survey data to their health and social care records. Approval was granted in May 2016. Approval was also sought and received for education data linkage from the Scottish Qualifications Authority (SQA).

To date, work to link HAGIS to NHS Health records (Hospital Episodes Statistics (HES) and GP Prescriptions) and education data has been undertaken. Work is ongoing to seek approval for access to dental, DWP and HMRC data.

Access permissions / conditions of use: Access to anonymised HAGIS data is to be made available through the G2G. To access anonymised HAGIS data linked to health and education records, researchers must apply to the PBPP and have had full information governance training in handling personal data.[23]

4.9 HAALSI Linkage:

The primary aim of HAALSI is to understand the drivers and consequences of HIV and non-communicable diseases among the ageing population in Agincourt, South Africa as little is understood about the nature of ageing in the region.[24] Reduced life expectancy caused by HIV has begun to reverse with the increasing availability of antiretroviral treatment (ART) and additional socioeconomic and health improvements.[25] However, increased life expectancies have been accompanied by a rise in chronic disease in adults throughout sub-Saharan Africa. Cardiovascular and metabolic diseases are now at unforeseen levels with incidence projected to more than double in the next 20 years. HAALSI explores the interrelationships between: physical & cognitive functioning, lifestyle risk factors, household income & expenditure, depression & mental health, social networks & family composition, HIV infection, and cardiometabolic disease.[26] The study has been designed to capture characteristics specific to the ageing process in rural South Africa, as well as harmonize with the HRS and sister studies on the G2G. HAALSI data is integrated with cause of death data from the INDEPTH Health and Demographic Surveillance System (HDSS) data at the MRC/Wits Agincourt research site.

Access permissions / conditions of use: Wave 1 Data are available through: The Harvard Dataverse Network, housed at the Institute for Quantitative Social Science (IQSS) at Harvard, The

[23] *ibid*
[24] https://g2aging.org/?section=study&studyid=37, accessed 04/09/20, https://haalsi.org/data, accessed 04/09/20.
[25] https://haalsi.org/data, accessed 04/09/20.
[26] https://haalsi.org/about, accessed 04/09/20.

Inter-university Consortium for Political and Social Research (ICPSR) at University of Michigan, INDEPTH Data Repository.

4.10 NICOLA Linkage:

NICOLA traces the health, lifestyles and financial situations of a representative sample of individuals aged 50 and over, as they grow older, monitoring how their circumstances change over a 10-year, period to inform policymaking in Northern Ireland. Currently, there is no mention of links to administrative data on either the NICOLA or G2G websites although work has taken place that links the survey to:

- NHS Health records
- Area Level Data – Index of Multiple Deprivation (IMD)
- House capital values – Rates (local authority tax)

Access permissions / conditions of use: Access to NICOLA data and/or biological samples requires completion of a research proposal form, to be submitted to nicola-research@qub.ac.uk for approval by the NICOLA Data Access Committee.

4.11 Other HRS Family Surveys:

The following HRS Family surveys make no mention of links to administrative data on either their own or G2G websites.[27]

4.11.1 LASI

LASI is designed to examine ageing and retirement among India's 45+ population, focusing on health, economic, and social well-being, for scientific research and evidence-based policy-making purposes. The study design has been informed by the HRS, including the same core contents, and is comparable to other health and retirement studies whilst taking account of the unique institutional and cultural characteristics of India.[28] It was launched under the aegis of the Indian Government's Ministry of Health and Family Welfare (MoHFW) with a view to informing the National Programme for Health Care of the Elderly (NPHCE) and Ministry of Social Justice and Empowerment (MoSJE) social and

[27] Three new health and ageing studies in their planning stages for Egypt (A Longitudinal Study of Egyptian Healthy Ageing (AL-SEHA)), Lebanon and Uruguay are likely to join the HRS Family in the future.
[28] https://g2aging.org/?section=study&studyid=36, accessed 14/04/2020.

economic security programmes.[29] The study is funded by MoHFW, MoSJE, US National Institute on Aging (NIA), UN Population Fund-India.

Access permissions / conditions of use: Wave 1 data are not yet available; the G2G website states that they were expected to come on stream in late 2019/early 2020. Only the LASI pilot (2010) dataset is available currently.

4.11.2 KLoSA

The main study objective of KLoSA is to inform policy making for an ageing population by providing the data needed 'to devise and implement effective social and economic policies'.[30] The study has an early qualifying age of 45+, reflecting the earlier transition to retirement in South Korea (Lee, 2007) and enabling examination of the relationship between economic activity in middle age compared to older age against a background of economic disruption caused by financial crises. At present, KLoSA has no administrative data linkages. Researchers are required to sign up to register to gain sight of the questionnaires and Coding Guide and, in the process of doing so, provide unnecessary personal information.

Access permissions / conditions of use: Registration with the Korea Employment Information Service (KEIS) is required, after which anonymized data, questionnaires, code books and other information can be downloaded.

4.11.3 CHARLS

CHARLS' main aim is to provide a high quality, nationally representative data resource on health and socioeconomic status for research into ageing. The study aims to examine how state-level policy variations are related to health and retirement behaviour, collecting information on local social and economic conditions, pension policy, health care and health insurance, and other social welfare policies (Lee, 2010).

Access permissions / conditions of use: Registration with the National School of Development, Peking University is required to access the publicly available, anonymized datasets. At present, these appear to be the four main waves, the harmonised data and the pilot data, which can be downloaded from the National School of Development after registration.

[29] https://lasi.hsph.harvard.edu/files/lasi/files/lasi_overview.pdf, accessed 16/04/2020.
[30] https://survey.keis.or.kr/eng/klosa/klosa01.jsp

4.11.4 IFLS

IFLS was originally designed as a household survey of individuals aged 26+ and first conducted in 1993. It provides over 21 years of longitudinal information on individuals, households and communities. It was redesigned at the fourth wave, in 2007-08, to collect data on health and retirement comparable to those in the HRS family of surveys.[31] A community survey of local conditions and resources is also conducted.

4.11.5 SAGE

SAGE was designed to examine patterns and dynamics of age-related changes in health and well-being in six low and middle-income countries: China, Ghana, India, Mexico, Russian Federation and South Africa, and to investigate socio-economic consequences of these health changes.[32] It is a longitudinal follow-up study of individuals aged 50+ years with a smaller comparison sample of adults aged 18–49 years, from nationally representative samples.[33] SAGE was developed as part of a Longitudinal Survey Programme, by the WHO Multi-Country Studies unit, to compile comprehensive longitudinal information on the health and well-being of adult populations and the ageing process (Lee, 2010).

Access permissions / conditions of use: Meta- and microdata are available through the online WHO Data Archive at no cost on completion/approval of a *Microdata request form*.

4.11.6 HART

The motivation for HART was to create a national longitudinal and household panel dataset to provide an evidence-base on ageing behaviour in Thailand to inform public policy and improve the quality of life and well-being of older adults.[34] HART is a multidisciplinary survey of the health, employment, socioeconomic status, and social / financial supports within family networks of individuals aged 45+ years.[35] The majority of the survey content was informed by HRS, KOSLA and CHARLS, with the aim

[31] *Ibid*.
[32] https://apps.who.int/healthinfo/systems/surveydata/index.php/catalog/212 accessed 12/07/2020.
[33] https://www.who.int/healthinfo/sage/en/ accessed 12/07/2020
[34] http://iic.nida.ac.th/main/?page_id=564, accessed 22/07/20.
[35] https://g2aging.org/?section=study&studyid=44, accessed 22/07/20.

of harmonizing HART with the HRS and sister studies, with additional questions incorporated to reflect the local context. [36]

Access permissions / conditions of use: The data are hosted by the NIDA Intelligence and Information Center (NIDA-IIC) and are available for academic users from the website of NIDA's Center for Aging Society Research (CASR).

4.11.7 ELSI

ELSI aims to investigate the social and biological determinants of ageing and the consequences for the individual and Brazilian society[37] with particular reference to health service use.[38] The Study explores the dynamic relationships between health, economic and social factors as individuals age and plan for and enter into retirement, aiming to inform public policy to improve well-being, health conditions and reduce inequalities (Kaiser, 2013). Study design is informed by ELSA to enable international comparisons through data harmonization (ibid).

Access permissions / conditions of use: Researchers can register to use publicly available data on the study website.

4.11.8 MARS

MARS aims to collect micro-level data on individuals aged 40+ throughout Malaysia to inform policy making and the formulation of the National Framework for Active and Healthy Ageing Malaysia.[39] Very limited information is currently available on the website which is indicative of the study being in the early stages of development. Listed module content suggests clear harmonization with the HRS family of studies on the G2G platform.

Access permissions / conditions of use: No information on data availability or access is currently on the Study website.

[36] http://iic.nida.ac.th/main/?page_id=564, accessed 03/08/20.
[37] https://g2aging.org/?section=study&studyid=42, accessed 07/09/20.
[38] http://elsi.cpqrr.fiocruz.br/en/about-elsi-brazil/aims/, accessed 07/09/20.
[39] http://ssrc.um.edu.my/malaysia-ageing-and-retirement-survey-mars, accessed 15/09/20.

4.12 Initial Observations:

Ten of the eighteen HRS family surveys were found to have administrative data linkage in the review of studies' websites: seven harmonised studies and three sister studies. As indicated above, in general, the HRS family of surveys do appear to be more advanced with respect to harmonised data but other longitudinal studies would seem to be more advance with respect to administrative data linkage. Three initial observations may be made from the review of the studies' websites. Firstly, where linkages exist, these are often intermittent and can become quite dated by the time the linked data are available. Linked datasets may then, in turn, become even more dated as a result of the access procedures that researchers face. In short, there appears to be a *longitudinal continuity of measurement* problem (CLOSER, 2019), whereby intermittent administrative data linkage may, in effect, create a number of cross sectional studies. Secondly, as might be expected, the development of linkage across the different studies is piecemeal and generally not uniform. Where the HRS family of surveys have been linked, they are linked to different types of data depending on data access conditions within different countries and study interests (as shown by SHARE linkage). Lastly, microlevel data linkage appears to be most successful and/or consistent where this has been built into the development and design of the survey from the beginning and when there is political or legal weight behind the establishment of the survey, as demonstrated by the HRS, SHARE and JSTAR studies. The challenges of administrative data linkage and the issues it raises will be explored further through a survey of HRS Family Surveys' research teams with the findings to be reported in the Second Interim Project Report.

Appendix

HRS Family Surveys with Harmonised Datasets								
Survey	**Start year**	**Waves**	**Latest Wave**	**Sample Size**	**Age eligibility**	**Sample Frame**	**Data Collection**	**Data Additional to Core**
Health & Retirement Survey (HRS)	1992/93	14	2018/19	30,000+	51	U.S. Metropolitan Statistical Areas (MSAs) and non-MSA counties Wave 4 refreshment from Centers for Medicare and Medicaid Services' database	Face-to-Face & Telephone	Biomarkers, cognitive assessment, genetic data
Mexican Health and Aging Study (MHAS)	2000/01	5	2018	15,000	50	Mexican National Employment and Occupation Survey (ENOE)	Face-to-Face	Biomarkers, cognitive assessment, genetic data
English Longitudinal Study of Ageing (ELSA)	2002/03	9	2018/19	9,500+	50	Health Survey for England (HSE)	Face-to-Face	Biomarkers, cognitive assessment, genetic data
Survey of Health, Ageing & Retirement in Europe (SHARE): 27 countries	2004/05	7	2016/17	30,700 baseline 140,000+ latest	50	Population registers where possible	Face-to-Face /Self-completion of common questionnaire[40]	Biomarkers, wellbeing measures
Costa Rican Longevity & Healthy Aging Study (CRELES): two cohorts	2004/05: cohort 1 2010: cohort 2	3 2	2009 2012	~3,000 ~3,000	60 55	Cohort 1: representative longitudinal survey Cohort 2: Population Census	Face-to-Face	Biomarkers, genetic data
Korean Longitudinal Study of Aging (KLoSA)	2006/07	6	2016/17	10,200 baseline 7,400 latest	45[41]	National Statistical Office Census	Face-to-Face	
Japanese Study on Aging and Retirement (JSTAR)	2006/07	4	2012/13	3,700 baseline 5,100 latest	50	Stratified, random sampling in 5 selected local government municipalities	Face-to-Face / Self-completion	
The Irish Longitudinal Study on Ageing (TILDA)	2010/11	4	2016	8,000	50	Postal address/geographic clusters, stratified by socio-economic group & geography	Face-to-Face / Self-completion	Biomarkers, accelerometery, brain imaging, oral health
China Health and Retirement Longitudinal Study (CHARLS)	2010/11	4	2015	17,700 baseline 21,000 latest	45[42]	conducted in ten provinces of the People's Republic of China	Face-to-Face	Biomarkers, cognitive assessment
Longitudinal Aging Study in India (LASI)	2016/17	1	2016/17	60,250	45	Census data for 30 states and 6 union territories	Face-to-Face	Biomarkers, diagnostic assessment of dementia (DAD)

Table A1: HRS Family of Surveys' Main Features

[40] Questionnaire modules may differ across countries to reflect institutional structure but it is still aimed to collect conceptually comparable data (Lee, 2010).

[41] Earlier qualifying age of 45 reflects the earlier transition to retirement in South Korea (Lee, 2007).

[42] Cut-off reflects earlier transition to retirement in Asian countries (https://g2aging.org/?section=study&studyid=4, accessed 11/05/20).

HRS Family Sister Surveys (no Harmonised Datasets)								
Survey	**Start year**	**Waves**	**Latest Wave**	**Sample Size**	**Age eligibility**	**Sample Frame**	**Data Collection**	**Data Additional to Core**
Indonesia Family Life Survey (IFLS)[43]	2007/08	5	2014/15	30,000+	unclear	Nationally representative sample frame across 13 out of 26 provinces	Face-to-Face	Biomarkers
Study on Global Ageing and Adult Health (SAGE): Six countries - China, Ghana, India, Mexico, South Africa, Russia	2007/2010	3	2018/19	40,000+ W1 ~40,000 W2[44] [45]	50	World Health Survey (WHS)	Face-to-Face	Biomarkers, wellbeing measures
Health, Aging, and Retirement in Thailand (HART)	2015	3	2019	5,600	45	5,600 representative households sampled in 13 provinces from 5 regions and Bangkok/vicinity	Face-to-Face	
Northern Ireland Cohort for the Longitudinal Study of Ageing (NICOLA)	2014	3	2019/20	8,500+	50	Health and Social Care (HSC) Business Service Organisation (BSO) database	Face-to-Face / Self-completion	Biomarkers
The Health and Aging Study in Africa (HAALSI)	2015	2	2019	5,059	40	Agincourt health and socio-demographic surveillance system (AHDSS), rural Mpumalanga province, South Africa	Face-to-Face	Biomarkers
The Brazilian Longitudinal Study of Aging (ELSI)	2015/16	1[46]	2015/16	9,412	50	2010 National Census conducted by Brazilian Institute of Geography and Statistics (IBGE)	Face-to-Face / Telephone	Biomarkers, wellbeing measures
Healthy Ageing in Scotland (HAGIS)	2016 (Pilot)	1	2016	1,057	50	Stratification by 11 Health Board areas, addresses randomly selected from official Postcode Address File (PAF) and screened by National Records of Scotland for eligible residents	Face-to-Face / Self-completion	
Malaysia Ageing and Retirement Survey (MARS)	2018/19	1	2018/19[47]	9,000 target	40	unclear	Face-to-Face	

Table A1 (Continued): HRS Family of Surveys' Main Features

[43] Originally established as a household panel survey, becomes a survey of ageing at Waves Four and Five.

[44] Wave 2 data does not include Russian Federation data (Lee, 2010).

[45] Wave 3 sample size to be confirmed.

[46] Follow-up waves planned according to G2G but no mention of this on ELSI website (https://g2aging.org/?section=study&studyid=42, accessed 07/09/20).

[47] Wave 2 indicated as taking place in 2020 on study website.

HRS Family Surveys with Harmonised Datasets		
Survey	**Collaborating Institutions**	**Study Website**
Health & Retirement Survey (HRS)	University of Michigan	https://hrs.isr.umich.edu/.
Mexican Health and Aging Study (MHAS)	University of Texas Medical Branch (UTMB), Instituto Nacional de Estadística, Geografia y Informática, University of Wisconsin, Instituto Nacional de Geriatría, Instituto Nacional de Salud Pública, UCLA	www.mhasweb.org
English Longitudinal Study of Ageing (ELSA)	University College London (UCL), Institute for Fiscal Studies (IFS), University of Manchester, NatCen Social Research, University of East Anglia (UEA)	www.elsa-project.ac.uk/
Survey of Health, Ageing & Retirement in Europe (SHARE)	Centrally coordinated at the Munich Centre for the Economics of Aging (MEA), Max Planck Institute for Social Law and Social Policy	http://www.share-project.org/home0.html
Costa Rican Longevity & Healthy Aging Study (CRELES)	University of California, Berkeley (UCB), University of Costa Rica	http://www.creles.berkeley.edu/
Korean Longitudinal Study of Aging (KLoSA)	The Korea Employment Information Service (KEIS)	https://survey.keis.or.kr/eng/klosa/klosa01.jsp
Japanese Study on Aging and Retirement (JSTAR)	The Research Institute of Economy, Trade and Industry (RIETI), Hitotsubashi University, University of Tokyo	https://www.rieti.go.jp/en/projects/jstar/
The Irish Longitudinal Study on Ageing (TILDA)	University College Dublin (UCD)	https://tilda.tcd.ie/
China Health and Retirement Longitudinal Study (CHARLS)	Peking University, Natural Science Foundation of China, Behavioral and Social Research Division of the National Institute on Aging (NIA), World Bank	http://charls.pku.edu.cn/index.html
Longitudinal Aging Study in India (LASI)	Harvard T H Chan School of Public Health, International Institute for Population Sciences (IIPS), Mumbai, University of Southern California (USC)	https://lasi.hsph.harvard.edu/

Table A2: HRS Family of Surveys' Collaborating Institutions & Study Websites

HRS Family Sister Surveys (no Harmonised Datasets)		
Survey	**Collaborating Institutions**	**Study Website**
Indonesia Family Life Survey (IFLS)	RAND, Center for Population and Policy Studies (CPPS) of the University of Gadjah Mada (Third and Fourth Waves, previously Lembaga Demografi, University of Indonesia), Survey METER, Yogyakarta, Indonesia	https://www.rand.org/well-being/social-and-behavioral-policy/data/FLS/IFLS.html
Study on Global Ageing and Adult Health (SAGE)	SAGE Waves 1-3 were funded by the U.S National Institute on Aging (NIA), WHO and, variously, countries' national governments' ministry/health departments. Future waves are to be conducted based on country interest and support with The World Health Survey Plus (WHS+) being a future vehicle for continuing SAGE[48]	https://www.who.int/healthinfo/sage/en/
Health, Aging, and Retirement in Thailand (HART)	Center for Aging Society Research, National Institute of Development Administration (NIDA). Funded by the National Research Commission of Thailand. Second pilot funded by National Higher Education Commission (NHEC). Wave 3, planned for 2019, to be funded by Thailand Research Fund (TRF)	http://iic.nida.ac.th/main/?page_id=564
Northern Ireland Cohort for the Longitudinal Study of Ageing (NICOLA)	Queens University Belfast + collaboration with inter-disciplinary panel of experts in various fields of ageing	http://www.qub.ac.uk/sites/NICOLA/AboutNICOLA/
The Health and Aging Study in Africa (HAALSI)	Harvard Center for Population and Development Studies, MRC/Wits Rural Public Health and Health Transitions Research Unit (Agincourt), The INDEPTH Network[49]	https://haalsi.org/about
The Brazilian Longitudinal Study of Aging (ELSI)	Oswaldo Cruz Foundation – Minas Gerais (FIOCRUZ-MG), Federal University of Minas Gerais (UFMG)	http://elsi.cpqrr.fiocruz.br/en/
Healthy Ageing in Scotland (HAGIS)	University of Stirling	http://www.hagis.scot/
Malaysia Ageing and Retirement Survey (MARS)	Social Wellbeing Research Centre (SWRC), University of Malaya, Institute of Social Research (ISR), University of Michigan, University of Tokyo, University of Southern California, RAND Center for the Study of Aging	http://ssrc.um.edu.my/malaysia-ageing-and-retirement-survey-mars/

Table A2 (Continued): HRS Family of Surveys' Collaborating Institutions & Study Websites

[48] https://www.who.int/healthinfo/sage/en/ accessed 12/07/2020

[49] A global network of health and demographic surveillance systems based in Ghana: http://www.indepth-network.org .

Domain	Dataset	Hertfordshire Cohort Study	1946 NSHD	1958 NCDS	1970 BCS	ALSPAC	Southampton Women's Study	Millennium Cohort	Understanding Society	Next Steps
Health	Civic Registry									
	Maternity Records									
	Secondary and Community Care									
	Primary Care									
	Social Care									
Education	Early Years									
	School									
	FE/HE									
	Children in Care									
Employment	Employment									
	Earnings									
	Benefits									
Criminality	Convinction/Cautions									
Spatial	Geo-Spatial									
	Environment Exposures									
	Neighbourhood, Services, Green Space									
Consumer	Credit Histories									
	Mobile Phones & Personal Sensors									
	Social Media									
Other	Participant Tracing									
	Home Energy									
	Vehicle Registration									
	Voter Registration									

Linkage Key: Established | In Development | Planned for the Future | No Current Plans | No Coverage

Table A3: UK ESRC Longitudinal Studies' Data Linkage Activities

Source: ESRC, 2017.

Bibliography

Beaumaster, S., Chien, S., Crosswell, A., Lin, A., Phillips, D., Valev, M., Wilkens, J., Yonter, V. & Lee, J. (2018), *Harmonized ELSA Documentation*, Version F (2002-2016), USC Dornsife Center for Economic & Social Research, Program on Global Aging, Health, and Policy

Bohensky, M. A., Jolley, D., Sundararajan, V., Evans, S., Pilcher, D. V., Scott, I. & Brand, C. A. (2010), Data Linkage: A powerful research tool with potential problems, *BMC Health Services Research*, 10:346, http://www.biomedcentral.com/1472-6963/10/346

Borsch-Supan, A., Brandt, M., Hunkler, C., Kneip, T., Korbmacher, J., Malter, F., Schaan, B., Stuck, S., Zuber, S. (2013), Data Resource Profile: The Survey of Health, Ageing and Retirement in Europe (SHARE), *International Journal of Epidemiology*, Vol 42, pp992–1001. https://g2aging.org/?section=overviews&study=share-aut, accessed 23/11/20.

Brett, C. E. & Deary, I. J. (2014), Realising health data linkage from a researcher's perspective, *Longitudinal and Life Course Studies*, Vol. 5, No. 3, pp283-298

CLOSER (2019), *Preparing for the future: Tackling the key challenges facing the UK's longitudinal population studies*, Conference report, CLOSER, March.

Coeli, C. M. (2015), The quality of data linkage needs more attention, Cad. Saúde Pública, Vol 31, No 7, http://dx.doi.org/10.1590/0102-311XED010715

Dattani, N., Hardelid, P., Davey, J. (2013), Accessing electronic administrative health data for research takes time, *Archives of Disease in Childhood*, Vol 98, No 5, pp391–392.

Douglas, E., Wilson, T. & Bell, D. (2015), Healthy Ageing in Scotland: The Pilot Survey, Stirling Management School, University of Stirling.

ESRC (2017), *Longitudinal Studies Strategic Review*, Report, ESRC

Fischer, A. J., Medagliab, J. D. & Jeronimus, B. F. . (2018), Lack of group-to-individual generalizability is a threat to human subjects research, *PNAS*, Vol. 115, No. 27, https://doi.org/10.1073/pnas.1711978115

Gasteen, A. S. (2019), *Subject Choice and Attainment: State Secondary Schools in Scotland*, PhD thesis, University of Stirling, Stirling.

Harron, K., Dibben, C., Boyd, J., Hjern, A., Azimaee, M., Barreto, M.L. & Goldstein, H. (2017), Challenges in administrative data linkage for research, *Big Data & Society*, https://doi.org/10.1177/2053951717745678

Harron, K., Gilbert, R., Cromwell, D. & van der Meulen, J. (2016). Linking Data for Mothers and Babies in De-Identified Electronic Health Data. *PLoS ONE*, 11(10): e0164667. https://doi.org/10.1371/journal.pone.0164667

Kaiser, A. (2013), A Review of Longitudinal Datasets on Ageing, *Population Ageing* Vol 6:5–27, Springer, DOI 10.1007/s12062-013-9082-3

Kho, M. E., Duffett, M., Willison, D. J., Cook, D.J. & Brouwers, M.C. (2009), Written informed consent and selection bias in observational studies using medical records: systematic review, *BMJ* 338:b866, March 12, doi: 10.1136/bmj.b866. PMID: 19282440; PMCID: PMC2769263

King, G. & Roberts, M. E. (2015), How Robust Standard Errors Expose Methodological Problems They Do Not Fix, and What to Do About It, *Political Analysis* (2015) Vol 23:159–179, doi:10.1093/pan/mpu015

Knies, G., Burton, J. & Sala, E. (2012), Consenting to health record linkage: evidence from a multi-purpose longitudinal survey of a general population, BMC health services research, 12,:52, https://doi.org/10.1186/1472-6963-12-52

Korbmacher J, Czaplicki C. (2013) Linking SHARE survey data with administrative records: first experiences from SHARE-Germany. In Malter F, Bo ¨rsch-Supan A (eds). SHARE Wave 4: Innovations & Methodology. Munich, Germnay: Munich Center for the Economics of Aging (MEA), Max-Planck-Institute for Social Law and Social Policy, 2013.

Lee, J.(2007), *Aging in South Korea: Findings from Korean longitudinal study*, Conference paper, International Union for the Scientific Study of Population (IUSSP) Seminar "Aging in Developing Countries: Building Bridges for Integrated Research Agendas", Santiago, 23-24 April), IUSSP, Paris.

Lee, J. (2010), *Datasets on Pension and Health – Data for Collecting and Sharing for Policy Design*, RAND Working Paper WR-814, November.

OECD (2016), Research Ethics and New Forms of Data for Social and Economic Research, *OECD Science, Technology and Industry Policy Papers*, No. 34, OECD Publishing, Paris, http://dx.doi.org/10.1787/5jln7vnpxs32-en

Wagner, J., Valliant, R., Hubbard, F. & Jiang, C. (2013), *Level-of-Effort Paradata and Nonresponse Adjustment Models for a National Face-to-Face* Survey, Ann Arbor, Michigan: Institute for Social Research, University of Michigan; 2013, https://hrs.isr.umich.edu/documentation/user-guides#18

Weir, D. R. & Ofstedal, M. (2016), *Evaluating Strategies for Reducing Field Costs in a Longitudinal Study*, Ann Arbor, MI: Institute for Social Research, University of Michigan; 2016, https://hrs.isr.umich.edu/documentation/user-guides#18