ESRC Centre for Population Change Report

Linking longitudinal studies of ageing with administrative data

Second interim report

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 authors would like to thank the research teams of the following studies for participating in the survey and for being so generous with their time:

Costa Rican Longevity & Healthy Aging Study (CRELES)

English Longitudinal Study of Ageing (ELSA)

The Brazilian Longitudinal Study of Aging (ELSI-Brazil): Professor Maria Fernanda Lima-Costa, Dr Cesar de Oliveira

The Health and Aging in Africa: A Longitudinal Study of an INDEPTH Community in South Africa (HAALSI): Dr Elyse Jennings, Mr David Kapaon, Dr Julia Rohr, Prof Stephen Tollman

Healthy Ageing in Scotland (HAGIS)

Health, Aging, and Retirement in Thailand (HART): Dr Dararatt Anantanasuwong

Health & Retirement Study (HRS)

Japanese Study on Aging & Retirement (JSTAR)

Longitudinal Aging Study in India (LASI): Ms Arunika Agarwal, Prof David Bloom, Prof Sekher

Malaysia Ageing and Retirement Survey (MARS): Dr Halimah Arwang, Prof Norma Mansor, Ms Nur Fakhrina Ab Rashid

Mexican Health and Aging Study (MHAS)

Northern Ireland Cohort for the Longitudinal Study of Ageing (NICOLA): Dr Frances Burns, Prof Frank Kee, Dr Charlotte Neville, Prof Dermot O'Reilly

Survey of Health, Ageing & Retirement in Europe (SHARE)

The Irish Longitudinal Study on Ageing (TILDA): Dr Ann Hever, Dr Mark Ward

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

© Anne Gasteen, Elaine Douglas, David Bell all rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission provided that full credit, including © notice, is given to the source.

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 1

1. Introduction 4

2. Current Linkage 8

2.1. Data Sources & Ownership 8

2.2. Linkage Coverage 8

2.3. Consent 13

2.4. Summary 13

3. Linkage Mechanisms & Representativeness 14

3.1 Linking Mechanisms & Processes 14

3.2 Representativeness & Consent 15

3.3 Summary 16

4. Application, Approval & Access Processes 18

4.1 Establishment of HRS Surveys' Linkages 18

4.2 Updating Linkages 20

4.3 Summary 21

5. Researcher Access to Linked Datasets 22

5.1 Safe Haven Access 22

5.2 Remote Secure Access 23

5.3 Third Party Access Restrictions 23

5.4 Summary 25

6. Use & Impact 26

6.1 Linked Dataset Use 26

6.2 Impact of Studies' Linked Outputs 29

6.3 Summary 31

7. Benefits and Challenges of Data Linkage 32

7.1 Benefits 32

7.1.1 Data Validation 32

7.1.2 Reduction of Survey Costs & Participant Burden 32

7.1.3 Increased Information and Spectrum of Research 33

7.1.4 Complementarity of Longitudinal Studies of Ageing & Administrative Data 35

7.2 Challenges 36

7.2.1 Bureaucratic Processes: Negotiating Agreements, Ethical & Legal Barriers 36

7.2.2 Time / Delays 37

7.2.3 Access (study & researcher) 39

7.2.4 Resources for Processing Linkage (Cleaning, Coding, Link Quality) 39

7.2.5 Consent & Trust (Sample Size) Security Confidentiality 40
7.2.6 Infrastructure 42
7.3 Summary 43
8. Studies with no Current Micro-level Administrative Data Linkage 44
8.1 Where linkage is being considered 44
8.2 Benefits & Challenges 46
8.2.1 Benefits 46
8.2.2 Challenges 47
8.3 Where linkage is not being considered 48
8.4 Summary 50
9. Concluding Observations & Recommendations 51
9.1 Concluding Observations 51
9.2 Recommendations 52
References 55
Appendix: Survey Questionnaire 56

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.

This second interim project report details the findings from a survey of 14 HRS family research teams conducted between June 2020 and June 2021. The survey focused on the benefits and challenges of linkage, examining application, approval and access processes, linkage mechanisms, linked data usage and impact.

Our assessment of the linkage landscape found that linked datasets were invariably owned by central or local government departments. Health and mortality records were the types of data most often linked. Linkage was intermittent, reflecting fragmented administrative data infrastructures with multiple data controllers. Most studies were reliant on respondent consent to undertake linkage.

Linkage mechanisms were determined primarily by national administrative data infrastructures. Linkage keys were usually dataset specific as data infrastructures tended to be fragmented and lacked unique ID numbers that would allow access to multiple datasets. Linkage was most often carried out using exact matching of dataset-specific ID numbers, rather than through a single linkage key that allows access to respondent records across multiple official administrative databases.

Establishing and renewing linkages was critically dependent on relationships with data controllers, bureaucratic processes, resources, and national administrative data architecture. The length of time taken to build relationships and navigate the application/approval process varied between two and ten years. The longer the time taken, the greater the risk of disruption from changes in regulations and/or regimes.

Ease of access to linked datasets varied across studies, ranging from the use of remote secure terminals, to visits to safe haven sites. The use of safe havens presents a barrier to researchers that can be costly and time consuming. The Covid-19 Pandemic appears to have initiated a move towards

greater use of remote secure terminal access. Linked data access protocols could vary within studies, with different access regimes in place for different data linkages.

Where linked datasets were available to third party researchers, recording of access to the data or outputs from this was patchy; either because it was not deemed necessary to track use, or it was not possible given data access arrangements. Data share agreements tended to include clauses requesting researchers to notify studies of outputs, but these were not enforceable.

Linked data usage was greatest, in general, where linkages had been in place longest and third-party access was not restricted. In the case of six studies – ELSA, HRS, JSTAR, NICOLA, SHARE, TILDA – linked data outputs were seen to have influenced either government policy, public debate and/or the future direction of research.

While the validation of survey data was the single most cited benefit of administrative data linkage (seven studies), eight of the ten linked studies (at time of interview) also cited the mutual complementarity of longitudinal study data and administrative data and the broadening of research possibilities that linkage enables.

HRS family linked data research outputs suggest that this is where the greatest value of linkage lies. Longitudinal study linkage with administrative data allows research questions to be explored that would not otherwise be possible with one type of data alone. It enables the *before* and *after* of life events or policy impacts and mitigations to be studied providing a better understanding of causality, risk factors and outcomes.

The main challenges that the studies faced when undertaking data linkage centred on the building of relationships with both data controllers and respondents, data protection issues, the availability of resources and fragmented administrative data infrastructures.

The development of trust was cited as key to building relationships with data controllers and, crucially, respondents. Given the reliance of most studies on respondent consent for linkage, if respondent trust was low this was likely to be reflected in low consent rates. Low consent rates had the potential to reduce linked data sample sizes to the extent that any marginal benefits were unlikely to outweigh resource costs.

Of the four studies aspiring to future micro-level data linkage at the time of interview, MARS was best placed for this with the integration of official databases and unique personal identification numbers for the population. Future ELSI-Brazil and LASI linkage should be able to utilize unique personal identification numbers in Brazil and India respectively. HART linkage would be taking place against a background of fragmented data infrastructures and no unique ID.

The MHAS experience demonstrated that, in some cases, the administrative data architecture may be simply too fragmented for linkage to be feasible.

Better informed government policy was seen as the key benefit of linkage by the currently unlinked studies while the building of relationships, political buy-in and respondent trust were cited as the main challenges.

- We make the following recommendations:

    - Studies should seek respondent consent at the outset of any linkage process, given that linkage agreements with data controllers often depend on having respondent consent and can take several years to negotiate.
    - Simultaneously, studies should explore whether, given their governing data protection legislation, respondent consent is strictly necessary, as the removal of the need for respondent consent simplifies linkage processes considerably.
    - The HRS family should pool and share linkage expertise in a formal forum, perhaps under the auspices of Gateway to Global Ageing (G2G), to avoid duplication of effort.
    - HRS Studies should monitor and document the outputs from linked data and their impacts to build support for the case for future linkage.
    - Where there is no system of identifying individuals uniquely and administrative data systems are fragmented, administrative bodies should seek to establish such a system in order to enable linkage across administrative datasets (e.g. employment, tax, benefit, health, education records).
    - Where administrative data architecture is fragmented, infrastructure to centrally manage administrative datasets, creating a single point of access, should be developed by administrative bodies.
    - The feasibility of linkage in real time, with studies able to link to multiple administrative databases simultaneously, on an as-needed, live basis, should be explored by administrative bodies.

1. Introduction

This is the second 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, identifying the extent and nature of linkage to date and the associated benefits and challenges.

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 in most countries. The increased use of digital records over the last twenty to thirty years has rendered such data accessible for research purposes. Individual administrative databases have been used to provide descriptive and/or trend analyses for field professionals and policymakers. Their linkage to longitudinal studies adds value to lifecourse analyses by providing more accurate insights into the interaction of socio-economic variables and the outcome trajectories of individuals. Therefore, for the worldwide HRS family, which follows the experiences of older people in many parts of the world, such linkage is potentially even more valuable. Linking administrative data to these studies allows, for example, the interaction between income, education and health to be followed over the life course and into older age to assess outcomes and potential needs. This provides invaluable intelligence for policymakers on current and future socio-economic challenges against the background of (generally) ageing populations, enabling better-informed evidence to aid the development of policy interventions.

The first Project report (Gasteen et al, 2022) provided an initial overview of current and planned HRS family survey linkage based on publicly available information and/or documentation on individual study websites and the G2G platform. It revealed that: (1) linkage often occurred intermittently and the additional information could become dated by the time the linked data became available, (2) studies' linkage was not uniform. It varied according to access conditions applied to administrative datasets and to study team research interests, (3) linkage was most successful and consistent where it was built into the development and design of a survey at the outset and had legal and/or political backing.

For this second Report, a survey of the 14 HRS family research teams was conducted from June 2020 to June 2021 (inclusive) to further investigate studies' administrative data linkages and provide a more detailed understanding of their content, use, challenges and value added. The survey

focused on: the benefits and costs of linkage, application, approval and access processes, linkage mechanisms, linked data usage and impact. The survey questionnaire is provided in the Appendix. The comprehensive coverage in the survey of the many issues surrounding administrative data linkage meant that it was not always possible for individual interviewees to fully answer all questions, as different aspects of linkage may be dealt with by multiple team members or partner organisations. Nevertheless, the responses provide a rounded view of the benefits and challenges of longitudinal study and administrative data linkage. Table 1 lists the participating studies. All ten of the HRS family surveys with data linkages took part plus four studies with no current linkage at the time of interview (ELSI-Brazil, HART, LASI, MARS). ELSI-Brazil has since undertaken individual level data linkage, bringing the total number of HRS family surveys with data linkages to eleven.[1] Ten of these eleven *linked* studies have individual level data linkage. Five of these (ELSA, HRS, NICOLA, TILDA, SHARE[2]) also had contextual data linkage; that is background geographic, socio-economic, environmental or social policy data that are not specific to the survey respondent. One study, MHAS, only had contextual data linkage. The three studies with no current linkages all aspired to develop data linkage in the future.

Individual & Contextual Linkages	**Individual Linkages**	**Contextual Linkages**	**No Current Linkage**
English Longitudinal Study of Ageing (ELSA)	Costa Rican Longevity & Healthy Aging Study (CRELES)	Mexican Health and Aging Study (MHAS)	Health, Aging, and Retirement in Thailand (HART)
Health & Retirement Study (HRS)	The Brazilian Longitudinal Study of Aging (ELSI-Brazil)		Longitudinal Aging Study in India (LASI)
Northern Ireland Cohort for the Longitudinal Study of Ageing (NICOLA)	The Health and Aging in Africa: A Longitudinal Study of an INDEPTH Community in South Africa (HAALSI)		Malaysia Ageing and Retirement Survey (MARS)
Survey of Health, Ageing & Retirement in Europe (SHARE)	Healthy Ageing in Scotland (HAGIS)		
The Irish Longitudinal Study on Ageing (TILDA)	Japanese Study on Aging & Retirement (JSTAR)		

Table 1: Participating Studies from the HRS Family

Several themes emerged from this survey that provide further insights into the observations made in the first Report. These were: the building of relationships with both data controllers and respondents, data protection issues, the availability of resources, administrative data infrastructure, marginal

[1] ELSI-Brazil linkage took place towards the end of 2021, after the time of interview; as a result, not as much substantive detail is available on this linkage as for other studies.
[2] http://www.share-project.org/special-data-sets/social-policy-archive-for-share-splash.html

benefit and cost/risk trade-offs. Figure 1 below summarises how these themes underpin the initial, first report findings and interact with each other influencing the ability and decisions of studies to pursue linkage. Data protection issues lie at the heart of linkage processes as a key concern of government and institutions, shaping trust and relationship building, and resource and infrastructure needs which ultimately determine whether data linkages are established.

The structure of this report is as follows: Section 2 - surveys' data sources, Section 3 - linkage mechanisms and representativeness, Section 4 - the establishment of linkages, Section 5 – Access, Section 6 – Linkage use and impact, Section 7 – benefits and challenges, while Section 8 examines the experiences of unlinked studies and Section 9 summarizes the survey findings and makes a number of recommendations.

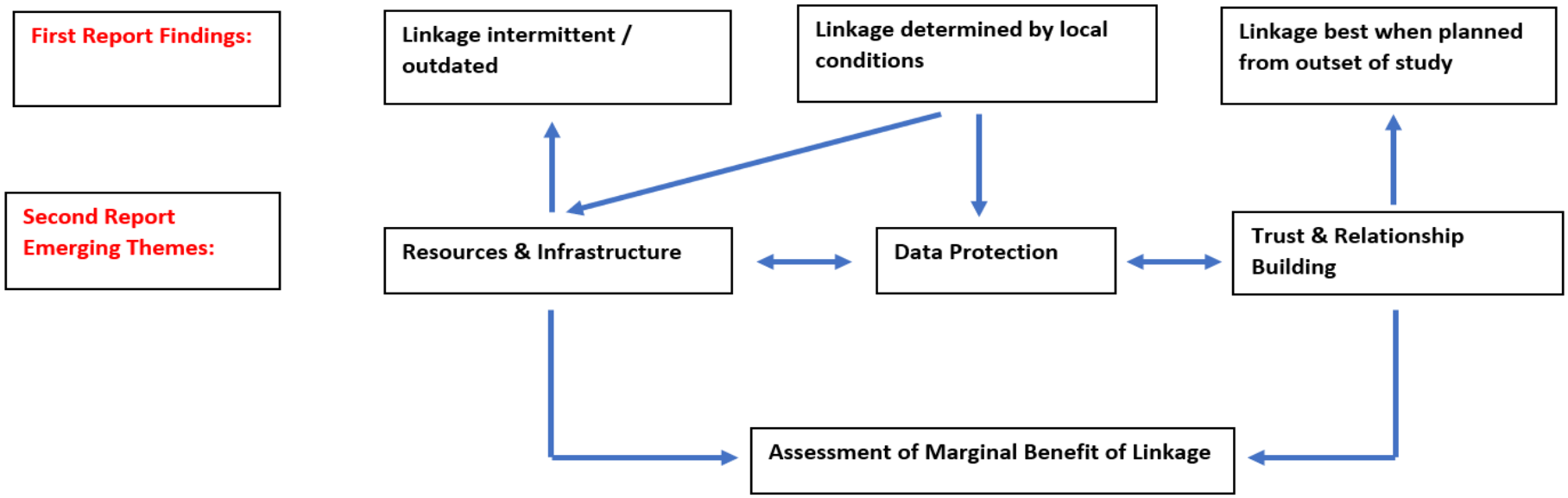

Figure 1: Interaction of Emerging Themes from Survey

2. Current Linkage

To provide an overview of current linkage, linked studies' data sources are listed in Table 2 together with the specific data controllers and time periods that the linked data covered. Current linkage is frequently intermittent (as identified in the first report). This reflects the degree of cohesion of administrative data infrastructure in different countries and the specific nature of linkage processes (as reported in Sections 3, 4 and 7). As might be expected, longer established studies tended to have a greater number of linkages.

2.1. Data Sources & Ownership

Most of the HRS family studies have health care data linkages; the exceptions to this are CRELES, ELSI-Brazil (at present), some SHARE studies[3] and MHAS which has no micro-level individual linkages. Just over half of the studies have linkage to death or mortality records. Only a few studies have economic data linkages: HRS, ELSA, some SHARE studies. The data controllers are invariably central or local government departments and agencies or national institutions such as, for example, the German Pension Fund (SHARE – Germany) or the NHS (ELSA, HAGIS and NICOLA).

2.2. Linkage Coverage

The time period that the surveys' administrative data linkages covered, and data completeness varied according to national institutional settings and structures. For instance, Northern Ireland has a well-integrated system whereby health and care records are linked, and individuals' health and care numbers can be linked to death records. In Ireland, whilst there is no unique patient identifier, work began in 2019 to link TILDA to individuals' General Practitioner (GP) records. Only GP records of practices that were on a new, emerging software system could be linked. In England, ELSA is not yet linked to primary care records as there is no national system with an all-encompassing, single GP register but different systems, with the best ones only thought to cover up to 30% of practices. Where there are separate but potentially overlapping health care systems, as in the US, a complete picture of health care utilization can only be obtained through linkage to all the different providers. For example, US forces veterans could be treated under either the Centers for Medicare & Medicaid

[3] SHARE is a multi-national survey, see Börsch-Supan et al, (2013) for further details.

Services (CMS) or the Veterans Affairs (VA) systems, requiring the HRS to be linked to both systems to provide a complete record of veteran respondents' health care.

For studies with multiple linkages, these varied by data type/owner reflecting the history/establishment of the linkage. For example, with ELSA the different data sources cover different time periods with no consistent overlap. Any research aiming to analyse individuals' health records alongside concurrent benefits and tax records would be limited to data for the three years 2006-2008.

Although administrative data might be linked to specific survey waves at the outset, the fact that studies' survey waves are linked for longitudinal analysis means that, effectively, all survey waves have administrative data linkage.

STUDY	DATA TYPE	DATA CONTROLLER [4]	TIME PERIOD
CRELES	Birth records Death records Voting records	Costa Rican government Supreme Electoral Tribunal	2005-2011 [No longer in field but continue to use date of death, carry out long-term follow-ups of individuals.]
ELSA	Mortality & Cancer Registration Records Hospital Episode Statistics (HES) National Insurance Contributions Benefits Tax Records Pollution data Energy Efficiency in houses	NHS NHS Department for Work & Pensions (DWP) DWP Her Majesty's Revenue & Customs (HMRC)	1998-2018 (Wave 9) 1948-2002 1977-2016 2006-2008 Forthcoming linkage Forthcoming linkage
ELSI-Brazil	Mortality Records	Ministry of Health, National Information System on Mortality	2015-2022
HAALSI	Census data: Agincourt Health & Demographic Surveillance System (AHDSS)	Wits Agincourt Research Unit	1992 onwards, annual data
HAGIS	Hospital Episode Statistics (HES) GP Prescriptions	National Health Service (NHS) Scotland NHS Scotland	4-5 years 9 years
HRS	Centers for Medicare & Medicaid Services (CMS) Veterans Affairs (VA) Social Security Administration (SSA): Earnings, Benefits Employer-provided pension plans National Death Index Census data, community surveys, air quality, crime, food access, medical resource distribution, long-term care use & facilities [5]	CMS VA released to HRS SSA HRS (derived variables based on pension plan information) US Federal Government Various	1991-2018 1999-2013 1992-2018 1992, 1998, 2004, 2010 2017 - latest HRS release Various

Table 2: Data Sources of Linked HRS Family Surveys

[4] Data controllers are individuals or organisations that determine the purpose and means of personal data processing. Data processors are individuals or organisations that process personal data on behalf of the controller. https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/key-definitions/controllers-and-processors/ accessed 08/11/2021.

[5] A full list of HRS linked datasets can be found at: https://hrs.isr.umich.edu/data-products/restricted-data/available-products

STUDY	DATA TYPE	DATA CONTROLLER [3]	TIME PERIOD
JSTAR	National Health Insurance (NHI) Long Term Care Insurance (LTCI) For 65+ New insurance scheme for medical services for 75+	Respective Municipal Authority (MA) MA Different local MA body from above, not yet agreed	2007 onwards 2007 onwards Forthcoming [not yet agreed]
MHAS	Mexican Population Census Mexican Public Health Sector Facilities Directory of Public Health Sector Facilities Census Literacy Rates by State of Birth Mortality Rates Administrative Records of Social Programmes: Progresa, Programa 70+, and Seguro Popular beneficiaries' enrolment Environmental Exposure and Data on Air Pollution	All publicly available data. Once linked, controlled by MHAS	2000, 2010 2002 2013 1910-1990 1998-2020 1997-2018 (Progresa) 2007-2018 (Programa 70+) 2001-2018 (Seguro Popular) 2007-2015
NICOLA	Area level data: Index of Multiple Deprivation Rateable house values NHS Central Register [forthcoming]	ONS Heath & Social Care Northern Ireland (HSCNI)	 Complete medical history but no birth records

Table 2: Data Sources of Linked HRS Family Surveys (continued)

STUDY	DATA TYPE	DATA CONTROLLER [3]	TIME PERIOD
SHARE			Varies by country, with some data available over life course, others not
Germany	German Pension Fund	German Pension Fund	Coverage from age 14 for Germany
Netherlands	Population registry, data on health, work, social security, income, wealth	Statistics Netherlands	
Denmark	Socio-demographics, employment, income, education	Statistics Denmark	
	National Patient Register (discharges, outpatient visits to somatic hospitals)	Danish Health Authority	
	Health Insurance Register (contacts to primary health care)		
	Pharmaceuticals Database (prescription medication)		
	Data on municipal services (home help, rehabilitation)		
	Cancer register (information on treated tumours since 1943)		
Austria	Demographics, employment (type, duration, income)	Social Security Institutions	
	Unemployment (duration, benefits)	Public employment service	
TILDA	Medical Card holders' data[6]	TILDA is data controller	From 50 years of age
	GP Records [in progress]	TILDA is data controller	Time spent with current GP
	Death Records	General Registry Office	2018
	Geo codes – environment / amenities		Since study respondent participation

Table 2: Data Sources of Linked HRS Family Surveys (continued)

[6] Medical Card eligibility is means tested.

2.3. Consent

For most studies, respondent consent was of critical importance for enabling the initial linkages and for obtaining updates, as it was often a requirement of data controllers. The exception to this was the MHAS study. MHAS linkage is to publicly available, contextual data only, requiring no respondent consent. For the other studies, the need to renew respondent consent varied.

Opt out consent for linkage to health records was obtained from all NICOLA respondents in February 2021. Respondent consent is required at each wave for TILDA; if individuals are eligible for a medical card, they are asked for permission to link to their medical card data. [7] For death record linkage, TILDA must reapply to the registrar for subsequent linkage, after first conducting a cross check on respondent deaths. Tilda does not have real time or automated data linkages. External data is sourced between TILDA data collection waves and this data is then linked at the individual level. SHARE linkage updates are also dependent on respondent consent, but this is valid until individuals either withdraw it or the end of the study. SHARE Germany has annual updates while for SHARE Denmark one update per wave is planned (previously there were no regular updates). Linkage for SHARE Netherlands is in real time. For JSTAR, agreements with individuals' municipal authorities (MAs) have to be renegotiated for linkage updates for each new wave, with access permissions being dependent on respondent consent.

2.4. Summary

The owners of potentially linkable datasets were invariably central or local government departments. Health and mortality records were the types of data most often linked, enabling research to include relationships between health and socio-economic variables. Linkage and its coverage were intermittent, reflecting fragmented administrative data infrastructures with multiple controllers. Most studies were reliant on respondent consent to undertake linkage, as this was usually a requirement of data controllers.

Administrative data infrastructure and respondent consent are discussed further in Sections, 3, 4 and 7.

[7] Protocol for updates of GP records not yet established as this is a new TILDA linkage.

3. Linkage Mechanisms & Representativeness

Linkage mechanisms and processes varied across studies, reflecting the different countries' institutional settings and administrative data infrastructures. Examining the ease with which studies can conduct linkage is important as, the more difficult this is, the more costly it is likely to be in terms of resources. Unique identification (ID) numbers, that provide a single linkage key with which to access respondent records across multiple official administrative databases, were rare. Linkage was most often carried out based on exact matching of dataset-specific ID numbers. This was supplemented by probabilistic matching in some cases. Most studies relied on respondent consent to undertake linkage.

3.1 Linking Mechanisms & Processes

In general, studies' respondents had different ID numbers for the various health and/or economic datasets. For instance, in the UK, ELSA and HAGIS linkage to health records is based on individuals' health care system numbers, while linkage to benefits and tax information requires respondents' National Insurance numbers. CRELES and SHARE Denmark were the exceptions; individuals in these countries have a unique ID number that is common across all official administrative datasets. ID card numbers were asked for as part of the CRELES' survey permissions to access births, deaths and voting data. CRELES is no longer in the field but continues to use date of death information and to carry out long-term follow-ups of individuals. SHARE Denmark uses respondents' personal ID/health card number - needed to access government services - as the linkage key.

National administrative data infrastructures combined with survey resources determined the nature and conduct of linkage processes. The HAGIS sample is drawn using the Scottish Community Health Index or CHI (equivalent to an NHS number in England), providing a unique ID number from the outset for health care linkage. Ad hoc interviewing of substitute respondents, however, required approximately 25% of the HAGIS sample to be matched by National Records Scotland (NRS), using probabilistic matching, to obtain CHI numbers. In TILDA, medical card numbers are used to link to prescription data, while unique GP ID numbers have been assigned to participants to enable the extraction of health records going forward. Death record linkage was problematic in Ireland as these were not held electronically but rather in a paper-based system. This linkage therefore required physical record searching and mapping which was only made possible by the funding of a dedicated post-doctoral post. HRS and SHARE Germany and Netherlands' linkage was based on social security

numbers with HRS using additional information (e.g., surname) for probabilistic matching checks. ELSA and NICOLA linkages are carried out by third party organisations. The independent social research agency NatCen provide the ELSA Team with the anonymised, cleaned data. NICOLA linkage is carried out by the Business Services Organisation (BSO) which provides businesses support functions and specialist professional services to the health and social care sector in Northern Ireland.[8] NICOLA linkage as based on exact matching where possible; usually in region of 50% to 66% of records with the remainder based on probabilistic matching using surname, address, date of birth and sex as the identifiers. The HAALSI sample is drawn from the AHDSS census data, enabling linkage to participants' household information and clinical records; AHDSS is linked to clinical records using probabilistic matching. The recent ELSI-Brazil linkage to mortality data used name, address, sex, age, mother's name, date of death and location of death to match participant records. For JSTAR linkage, copies of signed consent forms with respondent names and addresses are submitted to the MAs to extract the raw records from their databases and attach JSTAR ids. The JSTAR team clean and link monthly data returns, aggregating these over a year to link to the survey. This process involves aggregating and rationalising individuals' different medical episodes and treatments. MHAS contextual data linkages are performed using community numerical identifiers, at municipality, state (Entidad), or locality (Localidad) level. The linkages are undertaken in partnership with the Mexican Statistical Bureau (INEGI) which conducts both MHAS and the Mexican Census and uses the same identifiers across the datasets making the linkage process relatively easy.

3.2 Representativeness & Consent

As discussed in the First Interim Report (Gasteen et al, 2022), where respondent consent is required, this may impact on survey sample size and representativeness, potentially reducing statistical power and introducing bias (Kho et al., 2009, Knies et al. 2012). Public trust in national governments and institutions emerged as key to obtaining consent to link to personal data. TILDA reported high levels of respondent trust with high consent rates to access Medical card data and GP records. For TILDA Wave 5, 89% of the respondents consented to linkage. ELSA reported high consent rates for health data linkage of 97-98% but these were significantly lower for economic data linkage with a consent rate of 74% for DWP data linkage at Wave 8. HAALSI respondent consent to linkage was generally high, in common with other studies this varied according to the type of data linkage.

[8] https://hscbusiness.hscni.net/index.htm, accessed 27/05/2021.

SHARE studies' linkage consent rates reduced linked data sample sizes but were not found to have affected their representativeness and there had been no revision of weights. For SHARE Germany, the linkage rate was 65-70%, while for SHARE Denmark and SHARE Netherlands this was 80-90%. Low consent rates for SHARE Italy, however, meant that the viability of planned future linkage for this particular SHARE study was questionable. For CRELES, where requesting respondents' identity card numbers was part of the informed consent, it was acknowledged that sample size problems could arise from refusal or unknown card numbers, but refusal was found to be very small. The most complicated consent issues concerned biomarkers as these data are particularly sensitive.

The JSTAR consent rate was high at 80% but there were two particular issues with respect to the linkage arising from the sampling frame. Firstly, individuals who are not retired in Japan, are covered by company insurance. This means that, as the age eligibility for JSTAR is 50 plus, there are no centralised administrative medical records for non-retired, middle-aged JSTAR respondents. Secondly, representativeness and weighting are ongoing issues for JSTAR. Its sampling frame is based on 10 cities in preference to using the national, probabilistic, representative sampling employed by other HRS family surveys. The cities were identified first, then probabilistic sampling was employed and subsequently weighted according to national distributions. Municipalities are treated as fixed effects for comparison with other HRS studies.

For the HRS, the CMS consent rate was in the region of 85-95%. The VA consent rate was 50% of living participants and no weights were revised for this sub-sample. TILDA applied weights to adjust for the fact that, in common with other such studies, respondents tended to be skewed towards those with higher education and from advantaged socio-economic backgrounds. MHAS contextual linkages were applied across all survey respondents' records, except for air pollutants' data collected by monitoring networks, and there was no revision of sample weights. None of the studies reported any use of novel statistical methods in their analyses of linked administrative data.

3.3 Summary

National administrative data infrastructures and to a lesser extent, data collection methods, determined the nature of studies' linkage mechanisms. Linkage keys were usually dataset specific. No use, for example, has been made of social media to form linkages. In general, data infrastructures tended to be fragmented, having evolved piecemeal over time without unique ID numbers that would have facilitated linkage by providing a common key to access multiple datasets. Reliance on

respondent consent meant that where public trust was low, linkage might not be considered viable because of consequent low consent rates.

4. Application, Approval & Access Processes

Establishing and updating administrative data linkages can be very resource and time intensive with application, approval and access processes often at odds with project and funding timelines (Brett & Deary, 2014, Dattani et al, 2013). The modal time for HRS family surveys to work through the various application, approval and access stages to establish linkage was seen to be approximately two years. The background work to reach the application stage, however, meant that linkages often took the best part of a decade to put in place. In addition to lengthy bureaucratic processes, political regime changes could be challenging.

4.1 Establishment of HRS Surveys' Linkages

HAGIS was unusual among the studies as the team had to make an initial application for access to administrative data to draw the sample; the study's sampling frame is based on an administrative data spine. Overall, the application/approval/access process took approximately two years: one year to develop the application and obtain approval followed by access a year later. The application comprised two separate parts: linkage to draw the sample frame (that is, everyone, regardless of being approached or taking part), subsequent linkage for the consented sample. Consent for individual records' linkage was obtained as part of the survey fieldwork. Once the survey was established, further applications for linkage to Hospital Episode Statistics and GP Prescription data were made (Table 2). The greatest challenge was said to be the time cost of the safe haven access.

NICOLA linkage took approximately two years in total. For the Health and Social Care data, it took 15 months to lay the groundwork with the Information Commissioner's Office (ICO), the University and the Business Services Organisation (BSO) Honest Broker services before the research application to use the administrative data was written. Approval took three-four months with the final approval to access stage taking place within a six-month period. GDPR was implemented in the middle of the groundwork period, necessitating a re-writing of documentation.

With its Medical Card data linkage**,** TILDA had controller rights, so no formal application process was necessary. For other administrative datasets, the application, approval, access time varied between different datasets and data controllers. There were a variety of challenges in relation to the Death Records' linkage which proved to be very time consuming. Among these were the multiplicity of causes of death (approximately 5,000 codes), data quality control (as cause of death

selection was error prone and death certificate information varies widely) and prospective/retrospective Ethics/Consent/GDPR issues. Generally, access to Geo codes data had been a matter of months. It was felt that GDPR has focused attention on roles and responsibilities with respect to data access, explicitly formalising processes. If data is to be processed, robust IT lock and secure servers must be in place, and it is necessary to know who is accessing and/or using data. It was thought that the formal GDPR guidelines may have made this a bit quicker. The major challenge for the Geo codes linkage was using the environmental exposure variables in a format that did not increase the risk of identity disclosure.

For ELSA, the establishment of linkages varied across datasets. Generally, linkage was thought to take a minimum of two years from start to finish, in terms of NatCen receiving the data files. It often took longer, however, for instance the 2018 data process started in 2014/15. After receipt of the files, there was an additional delay of some two-three months while NatCen carried out the actual data linkage process.

Establishing linkage for SHARE studies was deemed to be easier because, with more countries taking part, it became possible to draw on the expertise developed on linkage projects in other SHARE countries. The major challenge to linkage was agreement with institutions. Data security is of paramount concern to data controllers, necessitating secure processes to work with the data, transfer the data and make sure respondents cannot be identified. Other projects provided a portfolio of different approaches that could be proposed to these institutions in relation to new linkages. For SHARE Germany, agreement with the German Pension Fund was reached before 2009 with linkage taking four-five years in total; final approval was given in 2009 with access established in 2011. SHARE Netherlands institutional agreements were reached in 2005 and respondent consent was asked for in 2013 (Wave 5). The time from final approval to access for this linkage was nine years (2011 – 2019). SHARE Denmark linkage received final approval in 2010 with access in 2017.

The application, approval, access timeline was not known for CRELES, with renewals being handled by the co-PI in Costa Rica. Access to administrative data for the MHAS team was instant as these were publicly available contextual data.

Linkage agreements for JSTAR could take from one month to one year to establish. The time from final approval to access was dependent on the time taken by the MA to draw down respondent records from their database. Reasons for any application rejections centred on confidentiality

concerns. The Confidential Data Protection Act 2004 was thought to have made things more difficult in this respect, but, in contrast, the new 2018 Confidential Data Protection Act should ease constraints for researchers. Bureaucratic processes and changes in political regime were identified as challenges. The former requiring patience while the latter could be problematic as it can take two years to re-establish a trusted, working relationship and convince those in office that the work is for research and not political purposes.

For HRS' linkages, the time from application to final approval could vary from approximately six months to two years. Processes could be quite idiosyncratic with specific concerns within/between agencies. Additionally, administration changes in Washington could lead to changes in the philosophy around data sharing which could cause delays. After final approval, access was relatively fast with the logistics of linking, data cleaning and checking of the merge usually carried out quite quickly. Initially, VA data access took about 10 years from the start of discussions to completion, as it was a start-stop process that involved multiple agencies. An example of the complexities around the VA linkage concerns data relating to alcohol/drug use, mental health issues and sickle cell disease. It is stipulated by the VA that these data cannot be accessed. HRS obtained living respondents' consent to access these restricted data, enabling the study to have this information for these individuals. After linkage, the alcohol/drug use, mental health issues, sickle cell disease data had to be removed for deceased respondents for whom no consent was in place. The subsequent data cleaning took quite a long time as processes had to be developed to do this.

HAALSI linkage at the US end took place on an informal, as needed basis, with requests for linkage made to the Wits Agincourt Research Unit. The time between the linkage request and receipt of the data varied from a few weeks to up to six months. This depended both on what variables were required and the urgency of the request. For instance, linkage to verify age would usually take just a couple of weeks. Sometimes small, non-urgent checks would take a couple of months. If the request was urgent, the Wits Agincourt Research Unit tried to process this as soon as possible.

4.2 Updating Linkages

In the broader, procedural context of updating linkage, resources, relationships, trust and changing bureaucratic structures and/or regimes were mentioned by other studies. CRELES has to seek formal reapproval with changes of government with the process dependent on the relationships and trust built up over time. Reapproval depends on the demonstration of due diligence in terms of data

protection with no breach of confidentiality. ELSA NHS linkages are periodically renegotiated against what is felt to be a continually changing bureaucratic landscape, with the establishment of NHS Digital being the latest example of this. In the UK, linkage negotiations with the DWP seemed to be particularly problematic across studies. Although DWP staff seem keen to encourage linkage, the process has been slow and minimally successful.

HRS CMS linkage updates require negotiation and agreement with the National Institute on Aging (NIA). The CMS linkage update for 2012-2018 has taken place over last two-three years, reflecting the complexities of update discussions. VA linkage updates requires NIA and HRS team interaction with the VA Data Office to reach a new agreement. The negotiation of such agreements is characterised by complex data security and use issues that change over time. It is believed that HRS is the only study that has received VA data. This was made possible because HRS specifically asked for individual consent to allow data to move outside the VA system. HRS has no current plans to update the VA data which cover the period 1999-2013 (Table 2) as, whilst it is desirable to keep this up-to-date, available resources have prevented this. It was suggested that the updating of this linkage needed 'champions' on both sides and that the funding and people to do the work had not been available to keep this linkage updated.

4.3 Summary

Establishing and updating linkages was critically dependent on relationships with data controllers, bureaucratic processes, resources, and national administrative data architecture. The length of time taken to build relationships and navigate the application/approval process, between two and ten years, meant that there was a risk of disruption from regime changes, as demonstrated by the NICOLA experience of the implementation of GDPR which necessitated the rewriting of the documentation.

5. Researcher Access to Linked Datasets

Once linkage is established, ease of access for researchers needs to be considered. The settings and protocols in place can deter or restrict the use of linked datasets, reducing potential research outputs. Ease of access to linked datasets varied across the studies as shown in Table 3. For some studies, researchers were able to access and/or download linked data remotely, through secure terminals. For others, access was much less convenient, involving physically visiting designated *safe haven* sites at appointed times. There were explicit third-party access restrictions to linked data for a few studies.

Safe Haven	Remote Secure Access	Third Party Restrictions
HAGIS	ELSA	ELSA
JSTAR	ELSI - Brazil	SHARE Denmark
NICOLA	HAALSI	
TILDA	HRS	
	MHAS	
	SHARE	

Table 3: Researcher Access to Linked Datasets

5.1 Safe Haven Access

Safe haven researcher access was used to varying degrees by HAGIS, NICOLA, JSTAR and TILDA. For HAGIS and NICOLA, researcher access to linked datasets was only provided through safe havens. While for JSTAR, access was usually on site as linked data required the highest level of security clearance. Online secure access protocols were possible but there was very limited capacity for this. Access to linked NICOLA data was provided either through the BSO safe setting or NICOLA's own safe setting. HAGIS linkage took place via secure access points at the Administrative Data Research Centre in Edinburgh. Researchers were required to: (1) be named on an application for use; (2) have a medical researcher as a guarantor; and (3) undergo two hours of online training (renewable every three years). Aside from the associated travel and time costs, the HAGIS team found the metadata to be poor and that there were data inconsistencies that required substantial cleaning. Secure remote access was permitted during the pandemic.

Researchers wishing to access TILDA linkages required permission from both data controllers to link and access the data; that is from TILDA for the TILDA data and the Health Board for the Prescriptions/Medicines' data. Prescriptions' Data linkage is carried out and accessed via a hot desk facility but with the Covid 19 pandemic, this is now changing to remote desk top access.

5.2 Remote Secure Access

SHARE linked data access requires application to and approval by SHARE and the provider institution. For SHARE Germany, application must be made to German Pension Fund; after approval, researchers receive a disk with the data or can access these through secure data transfer. For both SHARE Netherlands and SHARE Denmark, after approval, researcher access is via remote secure server. Austrian linkage is still in progress with the access routes yet to be determined.

The HRS' CMS Linked data are restricted access (highest level of security) and only available via secure remote access or under licence. Access involves a three-stage application process to: HRS, MedRIC who create and supply the data, and the Research Data Assistance Center (ResDAC) who authorise their use. The NIA also require CMS access to be signed off. Confidentiality & Security are paramount concerns for all sides: HRS, NIA, CMS. Linked VA data are also restricted access but this is via application to HRS only, though this is a simpler process.

Researcher access to linked MHAS data took approximately two months from application to approval with a further week for access. Once an application for restricted access MHAS (linked) data has been approved, researchers are able to download datasets using a one-time only link. Users of the linked data have to sign a 'restricted use agreement', whereby they agree not to try to identify MHAS respondents' municipalities (to protect identity) and not to distribute the linked dataset. Annual renewal of data agreements is required for continued use and confirmation of data deletion upon project completion.

HAALSI researchers receive files linking HAALSI to administrative data via encrypted email but are required to use a secure setting at Harvard for accessing sensitive biometric data. For CRELES, no specific application is needed to access the linked data since only the date of death is appended to the publicly available dataset. If researchers wanted the exact date of death, they must request this from the CRELES team giving justification.

5.3 Third Party Access Restrictions

In principle, there were no barriers to third party access to linked data for the HRS, HAGIS, NICOLA JSTAR, MHAS or CRELES. The HRS data access philosophy was to be as open and accessible as possible. JSTAR has a data control committee that approves third-party applications; to date there have been

no rejections. Similar to TILDA, access to the NICOLA linked data was contingent on an alignment of interests to the study team.

Third party researcher access to TILDA linked data was dependent on permissions from the data controllers. Applications are checked for feasibility and duplication with already approved and funded projects (as these are often independently and separately funded). Establishing remote desktop access for third party users was being considered because of the pandemic. It was expected that this would increase the number of users. Third party researchers are unable to access the Death Record linkage because of GDPR constraints on this type of data.

Third-party researcher access for HAALSI linked data was potentially problematic. HAALSI linkage requires direct interaction between the HAALSI team in the US and the team in South Africa. Any third party has to request that these linkages be made. Access is likely to take longer than it does for the core team, as a third party would need to apply for special data access and ethical agreements through the Wits Agincourt Research Unit.

There were explicit third-party restrictions in place for some SHARE studies and ELSA. Only researchers based in Danish institutions were permitted to access the respective SHARE study linked data. SHARE Germany and Netherlands linked data were available to researchers based abroad.

Researcher access to ELSA linked data is via UK Data Services and has become problematic over the last few years. Until 2015, third party researchers were able to access the linked data through UK data services. With the establishment of NHS Digital, this opportunity was closed. Only the core ELSA team can now access these data and no onward sharing is permitted. ELSA is actively trying to resolve the third-party access issue. NHS Digital also require retrospective deletion of previously acquired linked datasets. These can be impossible for ELSA to trace; individual researcher downloads could be identified through UK Data Services but not copied data files and additional users. This loss of third-party access has taken place against a background of new constraints coming into place, prior to the introduction of GDPR, in response to concerns about national data being lost. ELSA funders expect the data to be made more widely available than just the core team making the loss of third-party access a major problem.

DWP data linkage has become more difficult also as a result of increased UK government caution with respect to data sharing. It was felt that, in general, there has been a trend towards

increasing restriction. This has resulted in data that were not previously regarded as needing a special licence or secure access, being reclassified as such because of concerns about the possibility of identification. Data on rare health conditions were cited as an example of data that used to be widely available but are no longer. These trends are linked to an attempt to comply even better with general data security. GDPR added to ELSA's problems with respect to consent, as respondents were in the study long before its implementation.

5.4 Summary

The use of safe havens to access linked data is a barrier to researchers that can be costly and time consuming. The Covid-19 Pandemic, however, appears to have initiated a move away from safe havens, towards remote secure access. Linked data access protocols varied within studies, with different access regimes in place for different data linkages. Third party access to linked data was explicitly restricted for only a few studies; for ELSA this was the result of the changing data protection attitudes and landscape. For other studies, whilst there were no explicit third-party researcher restrictions, application processes could be complex.

6. Use & Impact

To apply for funding for linkage and/or advocate for investment in administrative data infrastructure, it is necessary to be able to demonstrate substantive research community use and impact of linked data outputs. As discussed in the previous section, not all linked datasets are available to researchers outside study teams. Where these are available, studies did not necessarily record access to the linked data or outputs from this. For some studies, this was because it was not deemed necessary to track use, for others it was not possible given the data access arrangements in place. Data share agreements tended to include clauses requesting researchers to notify studies of resulting outputs, but these were not enforceable. Hence, studies that kept records of data outputs on this basis could not guarantee their accuracy. Six of the ten studies with data linkage were seen to have produced impactful outputs from this, either influencing public debate and/or government policy.

6.1 Linked Dataset Use

Table 4 summarises study interviewees' responses on linked dataset use and outputs. The recording of access to linked datasets and resulting outputs from this was patchy; either because it was not deemed necessary to track use, or it was not possible given data access arrangements. The *Users* column indicates whether data is for use of the study team only. The *Outputs* column gives estimated outputs where interviewees felt able to do so and indicates those studies that did not track outputs.

Study	First Available	No. of Users	Users	No of Outputs
CRELES	2005		Not tracked	Not tracked
ELSA			Not tracked	Not tracked
HAALSI	Wave 1	3-5	HAALSI Team	
HAGIS	2017	3	HAGIS Team	3
HRS	1998/99 CMS 2015 VA SSA		HRS Team + External	
JSTAR	National Health Insurance Long Term Care Insurance	3-4 Level 3 >300 Level 2		
MHAS	2004	7-12		
NICOLA	2018	>5	NICOLA Team TILDA	Not tracked
SHARE Germany Netherlands Denmark	 2011 2019 2017	~ 80		Not tracked 30
TILDA	2012 Medical records/ prescription data & Geo Codes 2018 Death Records			

Table 4: Researcher Use of Linked Datasets

HAGIS linked datasets first became available in 2017. To date, only the HAGIS team have used these. There have been three direct outputs. Topics covered were smoking, loneliness and sample representativeness. At the time of interview, NICOLA linked data had only been available since 2018. Less than five researchers/teams had used it since then. Three to five pieces of work had been undertaken using the linked data covering: Representativeness, Dementia, Deaths/HES Cancer Registry linkage. This was expected to rise to five to ten outputs over the next year. There had been international collaboration with TILDA using linkage and interest in use had been expressed by the Commissioner's Office, Victims' Commissioner (Older Persons' Forum) and the Institute of Public Health.

TILDA medical records/prescription data linkage first became available circa 2012 with Wave 1 or early Wave 2 as did the Geo codes data. This prescription data was a planned link from the outset as TILDA was the data controller. The medical records linked data have been used extensively by researchers at the Royal College of Surgeons - who the lead on pharmacy/medication research - and by lots of students. Death Records first became available in 2018. There is a bureaucratic lag of several months in the availability of full death records. Death records' linkage enables mortality and co-morbidity to be investigated. This linkage is an ongoing project, with the second round of linkage taking place in the first quarter of 2022 and plans to regularise the process so that linkage can be undertaken on a regular basis. Currently, the death records' linked data are only available to the TILDA Team, but it is envisaged that they will be made available to external researchers at some future date. The GP record linkage is not currently available as this project is still in progress.

For ELSA, the Mortality Statistics are the most commonly used data linkage on the health side. The Study keeps a record of publications that it is informed of, making it possible, in theory, to identify outputs that used linked data for analyses of death. This does not accurately measure use, however, which is likely to be wider. ELSA does not formally record usage as the data are downloaded from NatCen. The Hospital Episode Statistics were first downloaded in 2018 and, therefore, were not available prior to the change to third party access conditions. Institute for Fiscal Studies (IFS) researchers were the main users of the economic data and it was not thought that there had been many other independent external users of this data without IFS collaborators. There has been third-party interest in using the linked data from the Department of Health and Social Care and the DWP, who have used linked ELSA data internally; for example, to carry out an analysis of disability benefits.

The IFS have tended to use linked ELSA data, rather than run their own, for government sponsored secondary analysis.

SHARE itself does not track users of linked data. Statistics on linked data use/download tend to be kept by the various institutional owners (who directly enable the download of their data) and are not shared. SHARE Germany linked data first became available in December 2011. By 2018, this linked data had been used by approximately 50 researchers/teams. There is no non-academic use of the data as they are only provided for research purposes. To date, there have been 20 outputs from the linked data covering pensions, health and the link between health and retirement age. SHARE Denmark linked data were first made available in 2017 and can only be accessed by researchers in Danish institutions (as mentioned above 5.3). SHARE Netherlands linked data became available in 2019 and can be accessed by researchers from other countries. SHARE Finland linked data should come on stream in the near future, while SHARE Austria linkage is still in progress.

The number of researchers using CRELES data is not tracked as most of these will be using the public access data. Only a small number of researchers have wanted access to the survival data; they need to make a request to CRELES for this but there is no formal process. There are, however, separate formal process for applications to access the genetic data as this is very sensitive. Outputs are not tracked. All outputs from CRELES data use the age variable which is verified through linkage. Only a small percentage of outputs use the mortality data.

The first MHAS linkage was created in 2004. Users are asked to notify MHAS of publications that come from MHAS/linked data. The study team check to see what registered MHAS users have published every few months, whether MHAS data were used and, if so, update their own records. Users sign an undertaking in the data share agreement to acknowledge the use of MHAS data and notify the Study of any related publications, but this cannot be enforced. Research areas using linked MHAS data included: socioeconomic conditions, mortality, health, health insurance coverage, environmental issues. Four researchers/teams used linked data in the first two years of its availability. Over the last two years, there have been six users: five domestic academic users and one international academic user.

JSTAR's highly confidential, linked administrative data with security access at level three (the highest level) were used by less than 5 researchers. All of those using the confidential linked data were academics, including a researcher at Stanford University. There were no outputs to date but

there is ongoing, international collaborative work with Stanford on future projections of long-term care/medical needs (to be published in the future). Linked JSTAR administrative data at security level two have been accessed by over 300 users. These were largely academic users with some government and third sector organisations (eg OECD), but no private sector organisations.

HRS CMS linked data have been available for more than 20 years, from the late 1990s. CMS usage runs into the hundreds. The data were used by both HRS Team and external researchers. CMS outputs were in the hundreds and, very possibly, thousands. Data users are requested to inform HRS of their outputs (publications/reports). The VA data have been available since 2015. VA related outputs numbered between 500 and 1000. The HRS online bibliography indicated that there were over 600 social security related outputs at the time of writing, but it is unclear if the SSA linkage was used in all of these.

HAALSI linked data had been available since the first wave as they were used for verification purposes. It was estimated that approximately three-five researchers have used the linked data to date; two-three using it in the last two years. There was no awareness of any third-party use. HAALSI does not directly record this but would be able to investigate the number of downloads of HAALSI linked data if this was needed for grant application purposes.

6.2 Impact of Studies' Linked Outputs

The NICOLA study has had an impact on the public debate about administrative data use through bottom-up advocacy. The team have presented and explained study research findings and how data are handled and protected to a panel of 100 NICOLA respondents. This has led to the realisation of the benefits of such work, with lay individuals being more comfortable with administrative data use, resulting in their willingness to support its use. In turn, this has made politicians more comfortable with the use of administrative data for research. The Team believed that there would be a COVID-19 effect on administrative data use processes, as the need to access data quickly in the pandemic has given impetus to speeding up access generally.

TILDA Medical Card data analysis has revealed the prevalence of over-prescribing in the Irish health care system. Changed legislation in 2013/14 specified that GPs should prescribe generic drugs where patients were happy to take these. As TILDA data straddled this legal change, the researchers were able to demonstrate that this was not happening; there had been no movement from the

prescription of high cost to generic drugs. The impact of this work led directly to changes in prescribing, away from high cost to low-cost generic drugs, reducing health care system costs. Additionally, it generated lots of interest in work on palliative care, homecare access, formal/informal care, nursing homes and causes of death. When the GP data come on stream, it is thought that this will increase understanding of the length of hospital stays and identify related issues. This will improve study data quality in this respect as self-reporting of hospital stays by older/unwell individuals can be quite inaccurate. Additionally, access to objective administrative health care data by Geo code, has enabled the identification of regional health care burdens and under/over-resourcing to inform health care planning. this.

Both ELSA and SHARE have contributed to domestic pension debates. A SHARE Germany paper, evaluating a supplementary pension benefit for those at risk of poverty, fed into the political debate on pension reform. This received a lot of publicity in the German media although the recommendations were not implemented. ELSA economic analysis has had a profound impact on pension debates. On the health side for ELSA, there has been work on loneliness, social isolation, health outcomes, mortality and wellbeing based on the Mortality data linkage. ELSA mortality studies have been widely cited in the literature and used in policy domains.

The original purpose of JSTAR was to inform the development of retirement policies and pensions. Now, it is to track and project future health care demand for Japan's ageing population. The study's impact lies in its work on future trends in frailty, dementia and the measurement of cognitive capacity and how this might affect the ability to work. JSTAR linkage to administrative government data has helped to shape policy focussing on how to extend retirement age in Japan.
HRS linked CMS data used for Dementia/Alzheimers' research in the US has likely influenced the debate. This work was presented in Congress, publicised by the NIA and has raised awareness of the long-term health and economic effects of severe illness. It is possible that this research has contributed to the increased funding for Dementia/Alzheimers' research. The work of Iwashyna and Prescott et al has focussed attention on the long-term impact of sepsis and other severe health shocks (see for example Iwashyna et al, 2010 and Prescott et al, 2016). It was emphasized that this type of research was not possible twenty years ago; it is linked datasets that have made this kind of work possible. Previously, clinicians could only focus on what happens in, for instance, the intensive care unit/hospital in terms of how quickly someone was discharged, whereas linkage has enabled analysis of pre-hospitalisation contexts and post-discharge impacts.

For CRELES there was no linked data policy impact per se as the linkage is used for data verification purposes (rather than adding information to individuals' records). For MHAS' contextual data linkage, it was not known whether subsequent work using this had had any particular policy impact.

HAGIS and HAALSI were both in the early stages of work/linkage that might lead to policy impacts. One of the five sub-projects within HAALSI aimed to use the Study data to inform policy related to taxes on sugar-sweetened beverages and limits on salt in processed foods. This work was likely to use the linked ADHSS data. HAGIS was still at the proof-of-concept, pilot stage and, as such there were no evident policy effects. The team have used ISD/Public Health Scotland cost data to consider whether smokers use more health resources. It is thought that this may have had a two-fold nascent impact viz: improving the administrative data quality and raising awareness amongst medical professionals to examine costs.

6.3 Summary

The responses captured in Table 4 suggest that linked data usage was greatest where linkages had been in place longest and third-party access was not restricted, as demonstrated by HRS and JSTAR use. For some studies, HAGIS and NICOLA for instance, linkage was quite recent (during the last three to four years) and usage to date was low as a result. However, there appears to have been relatively limited use of MHAS linked data given the early establishment of linkage for this study; this perhaps underpins the MHAS Team's view that researcher awareness of the potential benefits of linked datasets needs to be raised (see 7.2.3). The use of linked data has enabled pre- and post-event analysis to be undertaken to address policy-orientated research questions, something that was not possible two decades ago. In some cases, this has provided both an evidence base for policy changes and a focus for the direction of future research. Linked data outputs or work for six studies (ELSA, HRS, JSTAR, NICOLA, SHARE, TILDA) were seen to have influenced either government policy, public debate and/or the future direction of research.

7. Benefits and Challenges of Data Linkage

The linkage of the HRS family of longitudinal studies with administrative data can enhance the value of lifecourse analyses. The inter-play between socio-economic variables (e.g. income, education, health) can be traced out over the course of individual lifetimes into older age outcomes and potential needs. This knowledge provides improved evidence for policy interventions, in the face of the challenges presented by ageing populations. However, administrative data linkage can be very resource intensive, making it important to examine the challenges studies face in order to fully evaluate its use.

7.1 Benefits

The broad benefits of administrative data linkage identified by the participating studies are summarized in Figure 2 and Table 5 below. The two main benefits were seen to lie in the verification of survey data and a broadening of the research spectrum, enabled by the complementarity of the two different types of data. Reduced monetary and time costs were cited also.

7.1.1 Data Validation

The majority of studies identified the verification of respondent data, for accuracy and representativeness, as a benefit. Linkage to administrative datasets could be used for methodological research, to check survey representativeness, for example demographic response rates and returns on income could be validated. Administrative linkage enabled self-reporting inaccuracy and poor recollection issues to be overcome, therein, improving the accuracy and robustness of study survey data. It could be used also for respondent follow-up (CRELES, HAGIS) and survival analyses (CRELES, ELSA, TILDA), with cause of death analysis offering the possibility of identifying preventative interventions, e.g. better home/community care or GP access (TILDA). It should be noted, however, that while data validation is regarded as a benefit of linkage, administrative data can be subject to inaccuracies from reporting/entry errors and incompleteness.

7.1.2 Reduction of Survey Costs & Participant Burden

Four studies cited the potential to reduce survey costs as a benefit. Access to routine administrative data was viewed as having the potential to drive down survey costs significantly. More detailed information could be collected at lower cost, while interview time and space could be freed up to

collect other data, enabling a wider range of issues to be investigated. Allied to this was the reduced burden of data collection on study respondents in terms of their time and recall. Data linkage reduced the need to ask respondents for detailed data on, for instance, health incidents, income or labour market information (ELSA). Future access to primary health care records should reduce the need for expensive nurse bio-marker visits to collect, for instance, cholesterol and blood pressure levels (ELSA).

7.1.3 Increased Information and Spectrum of Research

Administrative data linkage was seen as a means of acquiring additional, more detailed information on study samples, adding a further longitudinal dimension and enabling the spectrum of research issues to be broadened through a two-way complementarity between panel survey data and administrative data. Multiple linkage was seen as contributing to the understanding of the ageing process in terms of its interconnected domains: health, economic & social circumstances (ELSA). This was viewed as invaluable in understanding the integrated nature of the ageing process, as analysis of individual domains was very limiting.

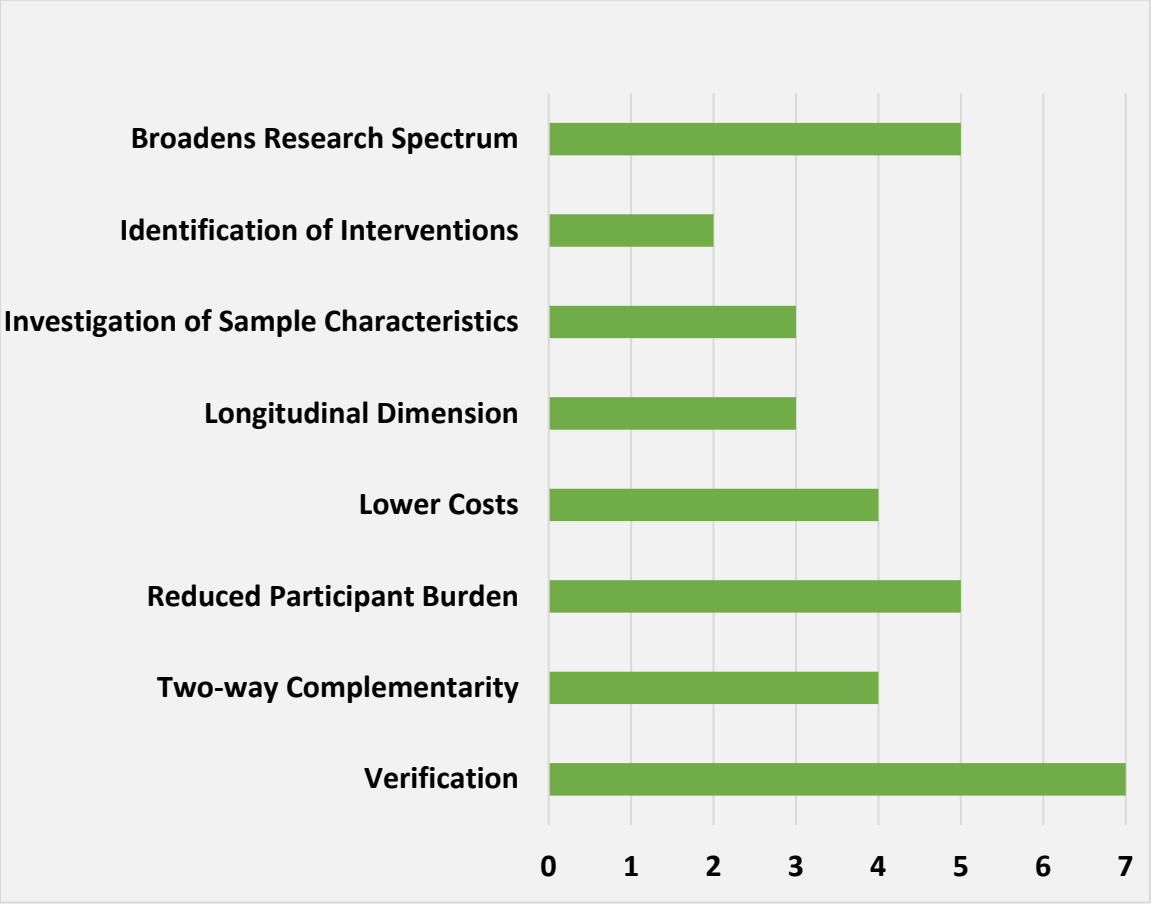

Figure 2: Benefits of Administrative Data Linkage

	Broadens Research Spectrum	Identification of Interventions	Investigation of Sample Characteristics [more info]	Longitudinal Dimension	Lower Costs	Reduced Participant Burden	Two-way Complementarity	Verification, Representativeness, Robustness of Data
CRELES								√
ELSA	√				√	√		√
HAALSI	√		√			√		√
HAGIS			√	√	√			√
HRS				√			√	√
JSTAR				√		√	√	
NICOLA	√				√	√		
MHAS	√							
SHARE	√	√	√				√	√
TILDA		√			√	√	√	√
TOTAL	5	2	3	3	4	5	4	7

Table 5: Benefits of Administrative Data Linkage

7.1.4 Complementarity of Longitudinal Studies of Ageing & Administrative Data

Several studies viewed linkage as creating a two-way complementarity between administrative and longitudinal survey datasets, enabling the exploration of issues that would not be possible with either data type on its own (HRS, TILDA, JSTAR, SHARE). The studies were unique panel databases that provided health, socio-economic and demographic context, enabling a greater depth of understanding of the issues than could be obtained from administrative databases alone. While administrative data contained accurate, retrospective information, ergo, linkage of the two types of data provided a combination of subjective and objective perspectives (SHARE, see also Börsch-Supan et al., 2018).

Specifically in the area of health, the value of administrative data linkage to longitudinal surveys could be seen in the linking of pre- and post- health event survey information on individual characteristics with health care utilization and/or diagnosis administrative data (HRS). Combining the two allowed for the examination of specific health conditions. More accurate information and detail on health episodes/shocks in terms of dates, timing, diagnosis and procedures were available from the administrative data. The administrative data complement survey data that is not collected annually by providing continuity in health utilization information. While the longitudinal survey data enabled a much better understanding of the long-term implications of health episodes/shocks in terms of physical and cognitive function and economic and/or family impacts through the provision of prior and post-episode data; something that was not possible with administrative data alone. The availability of prior data on health care status and socio-economic circumstances from the longitudinal surveys enabled health risk identification, providing a better sense of the trajectories both into and after a health shock.

The long-term implications of health problems among older adults can only be studied properly through a combination of longitudinal survey data and administrative health data (HRS). The linked datasets made it possible for clinicians to undertake *long-term impact* research; whereas 20 years ago, it would only have been possible to focus on, for instance, patients' periods of time in intensive care units (HRS). Linkage had also enabled subsequent work on the socio-economic implications for family members/carers of older individuals after health shocks (for example, having to move closer to an older relative that has had a health shock or reducing working hours to look after them). Demonstrating these increased possibilities for research was felt to build the case for further investment, while contributing to the general knowledge base of data linkage (NICOLA).

7.2 Challenges

The main challenges of administrative data linkage identified by the studies are summarized in Figure 3 and Table 6. Bureaucratic processes, resources and issues around consent and trust were cited by the majority of the studies. Time delays were mostly a by-product of these challenges; half of the studies explicitly referred to this as a problem. As might be expected, time to active linkage was shorter, the more established or closer the relationship between studies and data controllers.

7.2.1 Bureaucratic Processes: Negotiating Agreements, Ethical & Legal Barriers

Negotiating data share agreements can be lengthy and complex, and can take the best part of a decade (as discussed above, Section 4), when the initial groundwork of establishing contact with institutional data controllers and building relationships and trust is taken into consideration (SHARE, CRELES, JSTAR). The initial starting point can be political with top level cabinet or local government approval required (JSTAR). Application and approval processes to access administrative data are inextricably bound up with ethical and legal data protection and security requirements. Data security and/or consent issues can make negotiations complex and long. The approval of multiple actors/agencies at different levels can be required to ensure that data are secure and used appropriately, identity is protected, and individuals' preferences are adhered to (HRS, CRELES, SHARE, JSTAR, TILDA). Agreement at the top level, does not guarantee success as there may be refusal lower down the bureaucracy (JSTAR). Initially, the CRELES study had hoped to link to medical records. This proved not to be possible as those in charge of the relevant agencies at the time were concerned about data protection issues and unwilling to undertake these linkages.

In turn, lengthy, complex bureaucratic processes can impose time and financial costs on researchers and universities. In the UK, it was felt that while government encouraged research using administrative data, the cumbersome and lengthy bureaucracy involved meant that, effectively, the costs were defrayed to universities (HAGIS). Delays caused by application/access processes raised the costs of using the data in terms of academic time. Funders' deadlines might not be met, requiring 'no cost' extensions for which universities ultimately bore some of the cost. Legal or ethical requirements could change in the middle of an application process as, for instance, occurred with NICOLA and the introduction of GDPR. PhD student study could be seriously delayed by lengthy application processes, and it was not recommended that they worked with linked data unless the linkage was already established (HAGIS, NICOLA).

7.2.2 Time / Delays

The time taken to establish linkages was explicitly mentioned by six of the studies: HAGIS, NICOLA, TILDA, SHARE, JSTAR, HAALSI. In general, a large amount of administrative time was needed to set up linkages, which for TILDA meant that there was a need for *creative thinking* as to how to get access to what the Team needed to know. The most straight forward TILDA linkage was the Medical Care data as this was an established dataset that the team had controller rights for. [9] It was believed that government development of the proposed central data access centre in Ireland would help enormously together with the introduction of individual health identifiers (IHI). For SHARE, the search for appropriate institutional data holders and subsequent setting up of agreements could take years with the need to fulfil study, institutional provider and GDPR requirements. In addition to the time taken to establish linkages, the JSTAR team reported post-approval delays of a year or more in the receipt of administrative data records from MAs.

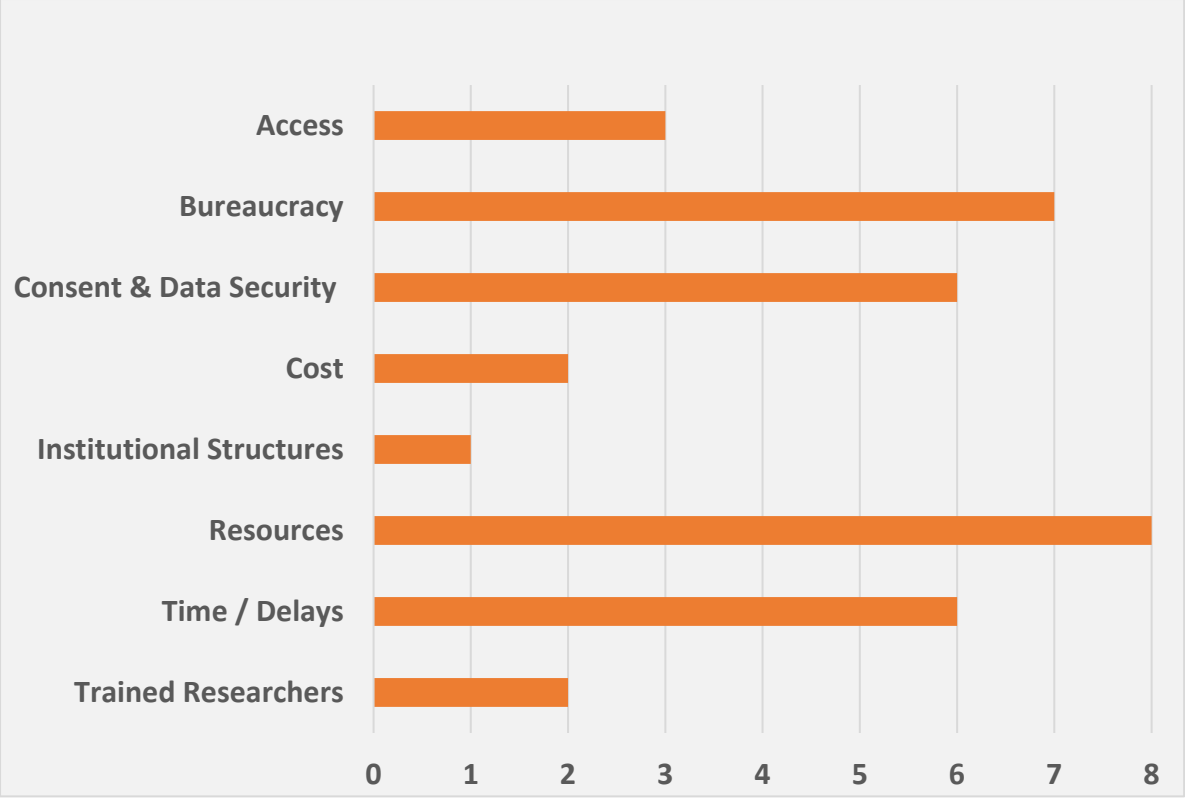

Figure 3: Challenges of Administrative Data Linkage

[9] Medical Card Prescription data comprised GMS medical card dispensing data only; long-term illness/drug payment schemes were excluded.

	Access (study & researcher)	Bureaucratic Processes: Ethical & Legal Barriers	Consent & Trust (Sample Size) Security Confidentiality	Cost	Institutional Structures	Resources for Processing Linkage (Cleaning, Coding, Link Quality)	Time / Delays	Trained Researchers
CRELES		√	√					
ELSA	√					√		
HAALSI							√	
HAGIS		√		√		√	√	
HRS		√	√			√		√
JSTAR	√	√	√			√	√	
NICOLA		√				√	√	√
MHAS			√			√		
SHARE	√	√	√			√	√	
TILDA		√	√	√	√	√	√	
TOTAL	3	7	6	2	1	8	6	2

Table 6: Challenges of Administrative Data Linkage

7.2.3 Access (study & researcher)

Access to administrative databases for studies and/or researchers, in the context of bureaucratic processes, was directly cited as problematic by three studies: ELSA, JSTAR, SHARE. For ELSA, data access (for both the study team and external researchers) was seen as the greatest challenge, restricting the potential research that could be carried out. Access for the wider research community was felt to be the biggest challenge for SHARE in terms of having to negotiate this across the different countries and institutions to make it as easy as possible. The accessibility of municipal authority databases in terms of their research readiness was an issue for JSTAR. The view of the MHAS Team was that awareness needed to be raised among researchers; users needed to be more familiar with the concept of linked data and recognise the potential for its use and the widening of the scope of research questions.

7.2.4 Resources for Processing Linkage (Cleaning, Coding, Link Quality)

The availability of resources for linkage processing was the most cited challenge (eight out of the ten linked studies: HAGIS, NICOLA, TILDA, ELSA, SHARE, MHAS, HRS, JSTAR). The ease of post-approval study access was dependent on administrative data infrastructures which could be fragmented and/or piecemeal. At present, TILDA's future GP linkage was unlikely to be comprehensive as a subset of practices were thought to be on the nascent software system (Data Sources and Ownership above, Section 2.2), reflecting the evolutionary, piecemeal, add-on development of the Irish health care system.[10] The TILDA Deaths Records linkage had been a manual process, requiring the scanning of Registrar General paper records, which would not have been possible without dedicated resource funding. The GP and death data linkages were both labour intensive processes incurring extra cost for the study. It was strongly felt that *"Studies should not have to solve health care data access and infrastructure problems but should be able to link into these data to support health care systems. Covid was believed to have shone a light on disparate infrastructures and an urgent need for timely, linked, holistic information"*.

Studies cited technical challenges (HRS), potential sample size, bias and reduced representativeness problems (SHARE), data cleaning, rendering and coding issues (TILDA, ELSA, JSTAR,

[10] As noted above, (Data Sources and Ownership), ELSA GP records' linkage was felt to be impractical at present for similar reasons.

SHARE) and a need to assess data quality (MHAS). Incomplete/missing data were an issue for TILDA's in-progress GP linkage with, for instance, vague GP practice addresses, no or incorrect GP names, relocated practices and retired doctors. Interviewer mistakes, with ID numbers recorded incorrectly in interviews, could be an issue (SHARE).

No specific technical challenges were identified for ELSA as the linkage work is carried out by NatCen, but there was a feeling that sometimes it was not completed as quickly as it could be. The coding/rendering of the data could be complex. For example, the coding of the multiplicity of Cancer Registration categories (incidence, mortality) poses problems of aggregation to form major categories. Finding the financial and human resources to work on complex datasets was an issue (HRS, HAGIS, NICOLA). There was a scarcity of trained researchers with the appropriate skills to conduct linked data analysis (NICOLA, HRS). For HAGIS, there was a lack of clarity as to who should resource/fund linkage; whether this should be surveys or data controllers. Linkage costs for the study were high, including fixed and recurrent variable costs for access, processing advice and output checking. Resource costs are sometimes built into grant applications but were not for HAGIS. HAGIS' inclusion under Scotland's Administrative Data Research Centre would have led to its linkage being resourced. In the event, the Team secured additional funding from the Nuffield Foundation for further HES and prescriptions' linkage.

7.2.5 Consent & Trust (Sample Size) Security Confidentiality

Issues around obtaining consent, developing relationships, and building trust, data security and confidentiality were identified explicitly as challenges by six of the studies (TILDA, SHARE, CRELES, MHAS, JSTAR, HRS). Studies emphasized the need to build relationships and trust with data controllers as the safeguarding of respondent confidentiality was of paramount concern (CRELES, JSTAR, HRS). Whilst MHAS did not link to individual, micro-level data, guaranteeing respondent confidentiality was at the core of its data use agreements.

Consent was regarded as a major challenge with most studies reliant on respondent consent for administrative record linkage. Low respondent consent rates can impact on sample size and representativeness, nullifying any benefits from linkage. TILDA had carried out a lot of work as to how best to approach obtaining consent, focussing on transparency, openness and the allaying of concerns. Trust in Ireland was seen to be high and was reflected in TILDA's high response rates to date. The respondent consent rate to link to GP records was unknown at present, with ethical approval having

been granted by the Trinity College Dublin Research Ethics' committee to seek participants' consent for this linkage prior to Wave 5. A Data Protection Impact Assessment (DPIA) had been carried out to ensure the protection of participants' data. SHARE Italy had sought respondent consent for linkage at Wave 8, but fieldwork was interrupted as a result of the Covid 19 pandemic, creating uncertainty with respect to the consent rate for linkage. Consent for linkage is to be sought at Wave 9.

Whilst most European studies rely on consent, this is not absolutely necessary under GDPR but it is possible that this may depend on individual country interpretations of the legislation. The need to fulfil both GDPR requirements and different countries' national legislative requirements and/or interpretations of GDPR was highlighted by the SHARE Team. For SHARE Netherlands, verbal consent for linkage was obtained at the time of interview. Whereas for SHARE Denmark, originally, consent was not regarded as necessary; respondents were informed and provided with an information sheet. After the implementation of GDPR, written consent was deemed to be required. With new linkage projects underway in a number of countries, SHARE studies are seeking respondent consent whilst simultaneously engaging in negotiation with data controllers. This was viewed as a pragmatic way to proceed because of the length of time between surveys and being in the field, with it being easier to have consent in place before/while institutional negotiations take place. SHARE aims to establish linkage for as many SHARE surveys as possible to maximise the mutual benefits of linked studies across countries. Linkage funding for SHARE studies was usually included in the national funding of surveys.

For the NICOLA Team, transparency was seen as key to participant stakeholder buy-in, which in turns leads to political buy-in. Advocacy among participants and the wider public, however, has the potential to be reduced with increased linkage of further datasets that have less obvious public benefit outcomes. It was felt that it was important for researchers to realise the existence of this trade-off and the consequent potential for reputational damage to surveys.

In the early stages of JSTAR, it was thought to be important to prioritise rapport and the building of trust with respondents and it was felt that consent to access death certificates (in the event) could not be asked for. The possibility of asking for consent for death certificate access was raised again five years ago, but there has been no agreement to date to pursue this and, currently, JSTAR is suspended due to financial reasons. The impact of Japan's Confidential Data Protection Act (2004) was greatest on JSTAR's death certificate linkage, resulting in MAs not allowing access to these. This linkage, with individuals' date and cause of death, is of fundamental interest to epidemiologists

but not economists (who tended to control the Study). Not to have death certificate linkage was perceived as a major weakness for JSTAR. Whilst the Study could access national records, probabilistic matching would have to be employed to enable linkage.

The ELSA Team is looking to add a large refresh sample in the near future and is exploring the use of the Family Resources Survey (FRS) for this purpose, as the Health Survey (from which the ELSA sample is drawn currently) may not be adequate in terms of numbers. This means that linkage consents will need to be examined to see if additional work is required to be done on these. The team are engaged in debates about the need for respondent consent for linkage as part of their rethinking as to how to do linkage.

7.2.6 Infrastructure

Another common theme that emerged was the fragmented and piecemeal nature of administrative data infrastructures that often had to be negotiated and overcome (HAGIS, TILDA). This made it crucial to carefully consider linkage options and complementarity. It was felt that establishing access to administrative data under one roof should make a difference in terms of application/access processes. The Irish Government is seeking to bring administrative datasets into a single hub (as mentioned previously) and a project is underway in Scotland to explore this also.

As part of infrastructure development, the need for a common ID number was identified. For the TILDA Team, the establishment of a single patient identifier would remove the need for multiple consents for access to, currently separate, health information datasets and systems. A patient identifier would supersede all of this and streamline the process. A unique identifier would be used to link across all health care sub-sectors with the result that consent to access patient records should only need to be asked for once.

The use of administrative data spines for drawing survey samples (as used by HAGIS) could also facilitate improved linkage infrastructure. Moving studies to administrative data spines, however, might be high cost. For instance, ELSA's planned refresh sample is, in part, a response to funder (NIA) concern about the lack of ethnic minority coverage; it is believed that using the FRS should provide a larger ethnic minority sample. Rather than moving to an administrative data spine, the FRS is preferred because it is a NatCen study, making the process easier. It was believed that switching to an administrative data spine would involve a lot more work and add to survey costs.

It could be perceived that the infrastructure for administrative linkage in the UK regressed with the ceasing of funding for the Metadac common pathway for data access for the Research Council's various longitudinal studies (that include ELSA). Metadac brought studies together, establishing a common process for allowing researchers to access more disclosive data. In principle, this could have included the development of administrative data linkage. Individual studies now need to have their own processes and may not have the resources. Where studies do have the resources, this could lead to a duplication of effort and a loss of shared knowledge. While studies opted out for specific ethical concerns with respect to their individual study data and a belief that there was a need for separate, study-specific sets of experts to make ethics' decisions, the ELSA view was that valuable economies of scale have been lost.

7.3 Summary

The main benefits of linkage were perceived to lie in the verification of survey data representativeness and the broadening of the research spectrum enabled by the complementarity of the two types of data. The linkage of longitudinal survey and administrative data enabled research into pre- and post-life events to be undertaken, providing a better understanding of causality, risk factors, outcomes, and insights into potential mitigations. Establishing data linkage was seen to take anywhere between two years at best and a decade at worst (Section 4). The time taken was dependent on the initial closeness of the relationships between studies and data controllers, the ease of negotiating ethical and legal barriers to establish data share agreements, and the nature of administrative data infrastructures. All of these factors could make linkage very resource intensive. Additionally, given the reliance of most studies on respondent consent for linkage, if respondent trust was low this was likely to be reflected in low consent rates for linkage. Low consent rates had the potential to reduce linked data sample sizes to the extent that any marginal benefits were unlikely to outweigh resource costs.

8. Studies with no Current Micro-level Administrative Data Linkage

This section explores the potential for micro-level administrative data linkage of the unlinked HRS family studies that took part in the survey and assesses where this might be implemented most easily. Of the four studies with no linkage at the time of interview: ELSI, HART, LASI, MARS, all indicated that the potential existed for micro-level linkage. The potential linkages identified by the four unlinked studies are shown in Table 7.

STUDY	DATA TYPE	DATA CONTROLLER
ELSI-Brazil	Mortality Data [11] Hospital Episode Statistics (HES)	Brazilian Ministry of Health
HART	Ministry of Public Health – cross-sectional health survey covering all ages National Statistics Ministry of Labor	Thai Government
LASI	Micro-level personal data	Indian Government
MARS	EKASIH scheme Ministry of Women, Family & Community Development Labour Force Survey Local government data – the Team are considering approaching local authorities to share MARS data to help inform local policy	Prime Minister's Dept Malaysian Government Ministry of Statistics Local Authorities

Table 7: Data Sources for Potential Linkage to HRS Family Surveys

At the time of this study ELSI linkage was at the planning stage, linkage to mortality records has since taken place. The other three studies wanted to pursue future linkage. LASI had no contextual linkages, but air quality and other data could be appended. Similar to the MHAS study, LASI used census data to validate the representativeness of the survey data. Whilst MHAS has contextual administrative data linkage (as discussed earlier), the study has no micro-level linkage although this has been explored previously.

8.1 Where linkage is being considered

Discussions were underway to connect ELSI-Brazil to a large health dataset based on SUS (Brazil's unified public health service) that was being developed by an ELSI-Brazil research collaborator. This meant that ELSI-Brazil's linkage would be to a dataset that had been cleaned already, rather than direct to SUS data. Individuals can use both public and private health care systems in Brazil and this linkage would not include private system records. Both waves of the main ELSI-Brazil survey (2015/16

[11] ELSI-Brazil linkage to mortality data took place towards the end of 2021, after the time of interview.

and 2019/20) and the five waves of the ELSI-Brazil Covid 19 Study (conducted by telephone interview during 2020/21) would be linked.

The existence of universal identification numbers, as part of the administrative data architecture, provides an important advantage when seeking to enable linkage. Multiple linkages can be undertaken with generally superior record matching accuracy. Individuals have universal identification numbers in Brazil, India and Malaysia but not in Thailand or Mexico, making any micro level linkage more problematic for the latter ageing studies.

The HART, LASI and MARS studies were at different stages in terms of laying the foundations for future linkage. When asked what was needed to enable linkage to take place (Table 8), for HART this required a broadening of the official mindset and a recognition of the benefits of data linkage covering the spectrum of activities the over life course of the population.

STUDY	What is needed to enable data linkage?
ELSI-Brazil	Keen research collaborators
HART	Government / bureaucratic awareness of benefits of linkage Champion with national influence
LASI	Development of a mechanism for data linkage that ensures data protection and security
MARS	Development of interface between databases - inter-operable application Data protection compliance – need to obtain respondent consent for linkage

Table 8: Linkage Enablers

The HART study is unique in Thailand, providing the only panel survey on ageing. The National Statistics Office (NSO) conducts cross-sectional surveys of older persons and employment while the Health Systems Research Institute (HSRI) conducts a cross-sectional survey on the Thai population's health (by health examination, covering the population from birth to 60 and over). The National Health Security Office (NHSO) holds the health data of people who are registered and covered by the National Health Security Fund. The Social Security Office (SSO) holds the data of the working population registered under the Social Security Fund. All of these government departments maintain separate databases for their own administrative purposes. There are no linkages between or among them. Recently, the Thai government's policy has focused on becoming a digital administration, the Ministry for Digital Economy and Society has been set up to facilitate the policy and utilize Big Data. The movement to set up the data linkage system among the relevant government agencies was still at an early stage. Moreover, it was felt that there was still a gap between research-based data (like HART) and administrative data that was yet to be bridged. The view was that a champion inside government, with national influence, was needed to progress a holistic approach with respect to data

collection and linkage. The need for linkage to be championed was implicit in the ELSI response also, citing a need for keen research collaborators to drive such projects as well as the availability of funding (ELSI is independently funded within Brazil and not reliant on NIA backing).

For MARS, enabling concerns were at the operational stage with the focus on the development of an interface between databases (rather than merging data) as the preferred option and, going forward, the need to obtain respondent consent to link to individuals' administrative records. The MARS Team already had a network of relationships in place, that they were continuing to build on, that would serve as a base to undertake linkage work in future.

MARS work has had exposure in government circles such as the National Advisory and Consultative Council of Senior Citizens, chaired by the Minister of Women, Family and Community Development, and the Implementation Coordination Unit ICU in the Prime Minister's Department. The MARS team intends to reach out to Department of Statistics for future linkage to the Labour Force Survey (LFS) which covers those aged 15-60. Hence MARS would provide valuable, later life, labour market insights as those working past 60 (more men than women) are not covered by the LFS.

The aim of the Social Wellbeing Research Centre (SWRC), in which MARS is based, is to further the case for stronger social protection and support government in this. More holistic social protection requires the inter-operability of databases across various government ministries and agencies. The Department of Statistics was working on this currently and, as such, the infrastructure for linkage existed.

8.2 Benefits & Challenges

Tables 9 and 10 list, respectively, the benefits and challenges of micro level data linkage identified by the unlinked studies. The focus of the challenges differed from that of the studies with established linkage, reflecting that these studies were at the beginning of the linkage process.

8.2.1 Benefits

The studies planning/aspiring to future micro level data linkage, all identified better informed government policy as a benefit (Table 9). For ELSI-Brazil, linkage-based research would inform the government about the nature of health service usage and facilitate analysis of the impact of the health

service on Brazil. ELSI-Brazil linkage would fulfil an increasing need for analysis of the ageing process, with research on the impact of policies to improve the evidence base for the development of future policy.

	Verification, Representativeness Robustness of Data	Lower Costs	Reduced Participant Burden	Investigation of Sample Characteristics [more info]	Longitudinal Dimension	Broadens Research Spectrum	Better informed government policy
ELSI							√
HART	√			√	√		√
LASI	√	√	√	√	√	√	√
MARS							√
TOTAL	2	1	1	2	2	1	4

Table 9: Benefits of Administrative Data Linkage

Better informed government policy, particularly with respect to regional differences was cited as a key benefit by HART. Currently, it was felt that regional differences were not reflected in a centralised government, 'one size fits all', approach to public policy in Thailand. Longitudinal panel data with linkage should give a new perspective to public policy formulation, informing policy better with respect to the background characteristics of individuals.

For the MARS team, with their emphasis on advocating for stronger social protection as Malaysia progresses to become a high-income nation, data linkage was seen as a means of enabling better delivery of social protection, financial and other support, including long-term care. It would enable policy preparedness for an ageing population and associated issues through the monitoring of health trajectories and health care needs. Other benefits identified were shorter survey instruments with no need to repeat questions on routine information, enabling a focus on the main areas of interest (LASI), data verification (HART, LASI) and greater investigation of sample characteristics with a longitudinal dimension (HART, LASI).

8.2.2 Challenges

The challenges identified by the unlinked studies were focused on two main categories (Table 10): bureaucratic processes rooted in ethical and legal considerations, and consent and data security issues. Understandably, no mention was made of access, cost, time delays or resources for processing linkage, as these are challenges that occur after linkage has been agreed.

The importance of government attitudes was cited by ELSI-Brazil – these were regarded as very positive towards data sharing in Brazil, and HART – where government was yet to be convinced of the benefits of linkage in Thailand. As above, Thai government bureaucracy was perceived as characterised by multifarious departments acting according to their separate legal mandates. In this respect, the lack of a single identification number across different official datasets was cited as a challenge also. Relevant government agencies were not aware of the nature of panel data and its benefits. When conversations had taken place on data sharing, these had not progressed well as a result of, it was thought, data protection concerns. All of the unlinked studies cited data protection and security issues as a challenge. The ELSI-Brazil HES linkage was thought likely to prove more difficult than Mortality linkage for this reason. LASI is currently discussing how and to what level data linkage can be undertaken without compromising the confidentiality of respondents. Data protection and security were of paramount importance for the LASI team as a loss of trust resulting from any data protection breach could have serious political and economic consequences.

	Bureaucratic Processes: Ethical & Legal Barriers	Consent & Trust (Sample Size) Security Confidentiality	Institutional Structures
ELSI-Brazil	√	√	
HART	√	√	√
LASI	√	√	
MARS		√	
TOTAL	3	4	1

Table 10: Challenges of Administrative Data Linkage

Trust was identified as an issue in obtaining respondents' consent for linkage by the MARS Team, with it being necessary to overcome individuals' distrust of government. The MARS Team perceived that they would have a further, specific problem with respect to obtaining consent, in that they would need to manage respondents' expectations and, therein, trust in the study. At Wave 1, the MARS Team passed on the name/address details of MARS' respondents in need, but not currently covered by any government support, to the appropriate government department. This raised expectations of receiving government assistance after the MARS' interview, which was not necessarily the case as this was dependent on official assessment of individuals' circumstances. This has become an issue for the Team in conducting Wave 2 and potentially for future waves.

8.3 Where linkage is not being considered

While MHAS has long-established contextual data linkages (as discussed above), the study had no plans for individual, micro-level data linkage and will not pursue this in the future. The Team are open

to other data linkages e.g. air pollution, particle matter data: PM10, PM2.5, and track other HRS surveys' linkage/use to see if they can do similar work.

Linkage to individual administrative data records was considered in the past. In 2012, the Team piloted the concept of mico-level data linkage when collecting biomarkers from a sub-sample of MHAS' respondents (approximately 2000 individuals). They explored whether or not it would be possible to obtain individuals' record numbers/IDs for data linkage (if allowed). Respondents were asked for their various ID numbers (there is no single ID number that covers more than one set of administrative data in Mexico) and their consent to link. Individuals' social security record number was asked for if they were affiliated to the Social Security Institute in Mexico (the formal labour market institute that provides health care in retirement to those who used to work in the formal labour market). Individuals affiliated to the [then] new retirement scheme (introduced in 1997) were asked for their federal numbers. Permission was asked of the Mexican retirement scheme for linkage but direct access to record files was denied and the team asked if the department would do the linkage.

During this pilot, the team encountered two main issues. Firstly, government agency agreement to link data was either not given or very slow. Secondly, respondent consent had the potential to significantly reduce sample size as the retirement schemes were not universal and very limited in coverage. When, eventually, permission was granted and a linked data file received, coverage was found to be very patchy. MHAS decided not to proceed with individual records' data linkage as it was considered too costly in terms of time and resources. Any benefit would have been very limited because of the effect on sample size (resulting from the limited coverage of such schemes). The end reward was too small for the obstacles that needed to be overcome to undertake the linkage. Therefore, MHAS had no plans to link to individual administrative data in the future.

The use of probabilistic matching algorithms was considered for verifying missing MHAS mortality records (based on details from relatives) by matching to official death certificates. The Team engaged with the Mexican Statistical Bureau who were willing to carry out this work. Subsequently, MHAS mortality records proved to be in line with statistical records and it was decided not to pursue this as it would not have meaningfully improved the survey records. Moreover, the Team did not regard probabilistic matching as a solution to obtaining a unique ID. The view was that there was more potential benefit in trying to get individual IDs established at the national level (rather than undertaking probabilistic matching exercises). The MHAS Team had been advocating the case for a new unique ID in public forums; that is that the establishment of such would make things easier for

both administrative and research purposes. A by-product of engaging in linkage work has been to demonstrate how difficult this is without a unique ID. Unique IDs would allow everything to be linked: census / school / employment / health insurance records. The influential Statistical Bureau of Mexico has been pushing also for the introduction of unique IDs.

The MHAS Team drew comparisons between micro-level data linkage in the US and the potential for this in Mexico. For MHAS, micro-level linkage would require engagement with a number of different agencies/institutions, each requiring different IDs, just to find, for example, retirement contributions from, for instance, social security or teachers' union records, depending on the schemes respondents were affiliated to. This fragmented provision, characterised by multiple systems with different IDs, stands in stark contrast to the US single social security system. Whilst it may not be possible to link all records in the latter, probabilistic matching could be undertaken to increase linkage. MHAS would have to engage with multiples agencies' data sources to do this, necessitating multiple permissions and data quality would be very variable. The use of probabilistic matching to increase record linkage would not be productive because names are not very good discriminators in Mexico, as there is a very limited range of surnames. This makes record matching for linkage much more problematic as names cannot be easily used. This very limited range of names is the reason other studies struggle to link Hispanics in US.

8.4 Summary

Of the four studies aspiring to future micro-level data linkage at the time of interview, MARS was best placed for this in terms of Malaysia's administrative data infrastructure with the ongoing integration of official databases combined with a unique ID number. Future ELSI-Brazil and LASI linkage would be able to benefit also from the existence of unique ID numbers in Brazil and India respectively. HART linkage would be taking place against a background of fragmented data infrastructures and no unique ID. Better informed government policy was seen as the key benefit of linkage by these three studies while the building of relationships, political buy-in and respondent trust were cited as the main challenges. The MHAS experience demonstrated that, in some cases, the administrative data architecture is simply too fragmented for linkage to be beneficial.

9. Concluding Observations & Recommendations

9.1 Concluding Observations

The single most cited benefit of administrative data linkage was the validation of survey data, an operational benefit; cited by seven out of the ten with established linkages at the time of interview. Eight studies, however, cited the complementarity of longitudinal study linkage with administrative data and the broadening of research possibilities. It is evident from the examples given by participating studies that this is where the greatest value of linkage lies. Longitudinal study linkage with administrative data allows research questions to be explored that would not otherwise be possible with one type of data alone. Linkage enables the *before* and *after* of policy impacts, events or shocks to be studied. This has been particularly evident in the field of health, where linkage allows clinicians to identify contributory risk factors to health shocks (longitudinal study data), assess the short-term efficacy of treatment (administrative data) and monitor the long-term impact of health shocks, treatment and recovery (longitudinal study data). The impact of this linkage can be seen, for example, in the discussions around Dementia/Alzheimer's in the US and the changes in the prescribing of drugs in Ireland (resulting from HRS and TILDA research respectively). The type of data most often linked were health and mortality records, leaving scope for further widening of the research agenda through socio-economic data linkages.

In general, the survey of HRS family research teams amplified the initial observations of the first project report. These were that: (1) linkage often occurred intermittently and the additional information could become dated by the time the linked data became available, (2) studies' linkage was not uniform and varied according to access conditions to different administrative datasets in the different countries and study team research interests, (3) linkage appeared to be most successful and consistent where it had been built into the development and design of a survey at the outset and had legal and/or political backing. A number of themes have emerged from this survey that provide further insights into these initial observations: the building of relationships with both data controllers and respondents, data protection issues, the availability of resources, administrative data infrastructure, marginal benefit and cost/risk trade-offs.

The development of trust was cited as key to building relationships with data controllers (usually government departments) and respondents. Some studies spoke of the need for *champions* in political circles to facilitate linkage to official data sources. It could be necessary also, to obtain agreement for linkage at different political or bureaucratic levels within or across organisations. The

length of time taken to build relationships and navigate application/approval processes could be anywhere from two to ten years, carrying a real risk of disruption from regime changes. Institutional data controllers' concerns were seen to focus on data protection issues and identification risk. Often, a prerequisite for the agreement of data controllers was respondent consent. Respondent consent was dependent on both the trust built up between study teams and participants, and broader cultural attitudes to government and officialdom. If low levels of trust resulted in low respondent consent rates, this had the potential to, at best, introduce or increase any sample selection bias or, at worst, reduce sample size to the point where any marginal benefit from linkage could be undermined fundamentally.

Administrative data linkage is highly resource intensive, with this being cited by the majority of studies with linkage as a challenge. Time and the availability of trained researchers to process linkages in terms of cleaning, coding and assessing linkage quality were seen to be problematic and could be compounded further where access was through a safe haven. The interaction of constrained resources with fragmented administrative data infrastructures and/or particular cultural features can negate any marginal benefit from linkage. This was the case with MHAS, where patchy coverage of multiple social insurance schemes together with a lack of a common ID number and insufficiently precise discriminators for probabilistic matching, had led the Team to conclude that there was no real benefit to be gained from individual record linkage.

Finally, the Covid 19 pandemic has further highlighted the value of longitudinal studies in general and demonstrated the need to link these with administrative data to inform policy in a holistic manner. Administrative data, on its own, will provide information on disease and treatment but will not offer an insight into the wider socio-economic and health effects of long Covid and associated illness or incapacity. The linkage of longitudinal studies with administrative data can provide such insights and enable better-informed policy making.

9.2 Recommendations

A number of insights in terms of best practice for administrative data linkage and the future facilitation of this can be gleaned from this survey of HRS' family research teams for studies and government administrations. Firstly, for studies, it is recommended that respondent consent for linkage should be sought as early as possible in the linkage process; if feasible, either prior to or at the same time as an initial approach to data controllers. This constitutes best practice on three fronts. Respondent consent is usually a prerequisite for data controllers. The time from an initial approach to actual

linkage(if successful) can be lengthy, sometimes as long as 10 years; therefore, where procedural permissions can be sought prior to or simultaneously with linkage negotiation, this should be done. Consent rates and their potential impact on sample size and representativeness will determine the feasibility of linkage in terms of the likely benefit to the study and whether or not to proceed with the process.

Secondly, it is recommended that, as well as seeking respondent consent at the beginning of exploring linkage, studies should investigate whether or not this is strictly necessary given the specific data protection and ethical requirements that they are bound by. If linkage without respondent consent is feasible, then clearly this would be the best option as it would negate any potential sample size or representativeness impact.

Thirdly, it is recommended that the HRS family formally pool and share linkage expertise. This tends to be done informally at present, for instance by TILDA (which shares its linkage expertise with other studies) and SHARE (where constituent studies' share their expertise), but it is felt that studies would benefit from having a formal forum for this, perhaps hosted by G2G. Further, it was evident that Studies do not tend to track outputs from linked data. Therefore, finally, for the HRS family, it is recommended that Studies should monitor and document the outputs from linked data and their impact in order to build and support the case for future linkage.

More streamlined administrative data infrastructure would improve access time and reduce resource costs of linkage. The following recommendations principally relate to government development of administrative data infrastructure for research and evidenced-based policy purposes. Firstly, where no single ID number exists and administrative data systems are fragmented, if governments are serious about utilising data linkage to inform policy making, then a single ID number needs to be established to enable linkage across administrative datasets: employment, tax, benefit, health, education records. Of course, this may encounter cultural push back and public trust could be an issue, as demonstrated in England with respect to the General Practice Data for Planning and Research scheme.[12] Clearly, data protection and privacy issues will remain and need to be resolved when allowing access for research purposes.

[12] Under the scheme, partially anonymised, GP health data for everyone in England are to be made available to researchers and companies for healthcare research and planning (https://www.theguardian.com/society/2021/aug/22/nhs-data-grab-on-hold-as-millions-opt-out, accessed 03/10/21).

Secondly, ideally, to overcome the fragmented nature of most systems and streamline access, infrastructure to centrally manage administrative datasets should be developed, as currently being explored by the Irish Government and in Scotland. Remote secure access systems should be established rather than safe havens, as the latter can be high cost in terms of researcher time.

Thirdly, linkage should be in real time. The Covid 19 pandemic has highlighted the need for this. Ideally studies should be able to link to *linked* administrative databases on an as-needed, live basis. This would overcome the problem of study linkages becoming out-of-date and remove the need for periodic manual updating by studies (although the need for *operational* updating would remain as new waves come on stream and participants' consent permissions change). As needed, live linkage to cross-linked databases would also streamline currently cumbersome and time-consuming access processes.

References

Börsch-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*. DOI: 10.1093/ije/dyt088

Börsch-Supan, A., Czaplicki, C., Friedel, S., Herold, I., Korbmacher, J. & Mika, T. (2018), SHARE-RV: Linked Data to Study Aging in Germany, in Winker, P., T. v. Büttner, R. Riphahn, W. Smolny, J. Wagner (eds.), *Jahrbücher für Nationalökonomie und Statistik*, Berlin, De Gruyter. DOI: 10.1515/jbnst-2018-0034

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

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.

Gasteen, A., Douglas, E., & Bell, D. (2022) Linking longitudinal studies of ageing with administrative data: First interim report, ESRC Centre for Population Change Report, University of Southampton.

Iwashyna, T. J., Ely, E. W., Smith, D. M. & Langa, K. M. (2010), Long-term Cognitive Impairment and Functional Disability Among Survivors of Severe Sepsis, *Journal of the American Medical Association (JAMA)*, Vol. 304, No. 16 pp1784-1794.

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

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

Prescott, H. C., Osterholzer, J. J., Langa, K. M., Angus, D. C. & Iwashyna, T. J. (2016), Late mortality after sepsis: propensity matched cohort study, *BMJ* 353:i2375

Appendix: Survey Questionnaire

Data Linkage of HRS Family Studies:

Survey of Principal Investigators/Study Teams

Thank you for agreeing to participate.

In preparation for the interview (via Microsoft Teams), please consider your answers to the set of questions that is most appropriate to your study.

1. Is <<survey name>> currently linked to any administrative data?

   That is, respondents' personal data (normally) requiring their consent or contextual data (e.g. employment rates) that does not require consent.

   If yes, please consider your answers to the following questions. **If no**, please go to Question 2

   A. What administrative data sources is <<survey name>> linked to?
      i. Who owns these data?
      ii. What time period do they cover?
      iii. What is the procedure for obtaining updates of these data?

   B. Which Waves of <<survey name>> are linked to these data?

   C. How is <<survey name>> linked to these data?
      i. What is the process for linkage to the different data sources?

   D. How has sample size / representativeness been affected by the linkage?
      i. Has it been necessary to compute new sample weights?
      ii. If so, how have these been revised?

   E. How do researchers access the linked data?

   F. When did the linked datasets first become available?

G. Approximately how many researchers/teams have used the linked datasets since they became available?
    i. How many researchers/teams have used, or applied to use, the linked datasets during the first two years of their availability?
    ii. How many researchers/teams are currently using, or have applied to use, the linked datasets in the last two years?
    iii. Do other researchers outside the <<survey name> team use these linked data sets?
    iv. If so, how many other users are there and in what sectors are they based? Please refer to the table below to answer.

	Domestic	International
Academia		
Public / Third Sector		
Private Sector (e.g. Consultancy)		

    v. Are there any barriers to third party access to the linked data?

H. Approximately how many outputs have there been that have used the linked data?
    i. What areas of research have these tended to cover?

I. What do you believe are the main benefits of this linkage to date?

J. How have your study's linked data research outputs influenced public debate and/or the policies of government (central/state/local), private business or the third sector?
    i. For whom might there have been an impact and in what way?
    ii. Have any mitigating interventions/strategies been developed as a result of this research?

K. What have been the main challenges or obstacles to linkage?

L. For each different linkage:
    i. How long was the time from application to final approval or rejection?
    ii. How long was the time from final approval to access?
    iii. Where the team had a rejection/s, in general, what were the reasons for this?
    iv. Were there any particular challenges or obstacles at these different stages of application/approval/access?

M. Do you consider that the statistical methods that you have used to analyse the linked data are novel and, if so, how?

N. Are there any other comments or observations that you would make about administrative data linkage?

2. Is it planned to link <<survey name>> to any administrative data?

If yes, please answers to the following questions. **If no**, please go to Question 3

A. What administrative data would <<survey name>> be linked to?
   i. Who owns these data?
   ii. What time period do they cover?
   iii. What is the procedure for obtaining updates of these data?

B. Which Waves of <<survey name>> will be linked to these data?
   iv. Is there any benefit to be gained by from linkage to more than one wave?

C. How will <<survey name>> be linked to these data?
   i. What will be the process for linkage to the different data sources?

D. What is needed to enable this data linkage?

E. How will researchers be able to access the linked data?
   i. Are there likely to be any barriers to third party access to the linked data?

F. What do you believe will be the main benefits of this linkage?

G. What potential impact do you think research outputs from this linkage might have? That is, for whom might there be an impact and in what way?

H. What do you see as the main challenges or obstacles to linkage?

I. Are there any other comments or observations that you would make about administrative data linkage?

3. Is there the potential to link <<survey name>> to administrative data?

   If yes, please consider your answers to the following questions. **If no**, please go to Question 4.

   A. What administrative data would you see <<survey name>> being linked to in the future?
      i. Would this linkage be retrospective, linking previous Waves to these data?

   B. How do you think the data could be linked to <<survey name>> ?

   C. What would be needed to enable data linkage?

   D. What do you think the main benefits of such linkage might be?

   E. What potential impact do you think research outputs from such linkage might have? For whom might there be an impact and in what way?

   F. What do you see as the main challenges or obstacles to linkage?

   G. Are there any other comments or observations that you would make about administrative data linkage?

4. Why do you think that the potential does not exist to link <<survey name>> to administrative data?

   A. Do you think that this could change in the future?
      i. If so, in what way?
      ii. Would this be desirable or useful?

   B. What would be needed to create the potential for linkage?

   C. Are there any other comments or observations that you would make about administrative data linkage?

Thank you for your time.