Longitudinal Studies Strategic Review: Annexes

Contents

Annex A: Outline of the review process, methodology and timetable

ESRC Office

October 2016

ESRC appointed the small, independent international Review Panel to undertake the Longitudinal Studies Review and a UK-based Steering Group.

The role of the Panel has been to shape and conduct the Review and its constituent elements, to assess the evidence gathered and to deliver a report with recommendations to ESRC.

The Steering Group comprised of members of ESRC Council and Capability Committee, plus individuals from the Medical Research Council, Wellcome Trust and GO-Science. This group was enlisted to provide strategic oversight and information to the Panel on the UK context.

October — November 2016

ESRC ran an initial open consultation via an online survey, which was widely promoted, seeking to engage a broad spectrum of stakeholders with an interest in longitudinal studies. The consultation received 637 valid responses from individuals in the UK (83 per cent) and internationally (17 per cent). Most respondents were from the academic sector (80 per cent) with the remainder spread across government, civil society and business sectors. Most respondents (81 per cent) had a background in economic and social research, with medical researchers accounting for 17 per cent.

Respondents identified key ongoing scientific priorities for longitudinal research: long-term effects of childhood and adult experience, demographic shifts and mobilities, health and wellbeing, equality and inequality, biosocial research and genomics, diversity and identity, ageing population.

Respondents also identified priority methodological and technological issues that longitudinal research needs to address: longitudinal study design, data collection, data handling and treatment, data analysis, data linkage.

An initial report on the consultation responses was produced along with ten more detailed and specific briefing papers.

9th & 10th January 2017

ESRC held an invitation-only two-day workshop at Nuffield College, University of Oxford, which sought to explore:

- What are the future scientific and policy needs for data to address the types of research questions for which longitudinal data has typically been used?
- How can these needs be met and what are key challenges to be addressed?

Experts from a range of academic, policy and professional communities (from around the UK and overseas) and members of the Panel and Steering Group were invited in order to reflect a wide range of key perspectives. Participants were asked to take a future-looking challenge-led approach and were encouraged to be critical, ambitious and creative.

A summary report of the workshop was published; this and more detailed information gained from the workshop helped inform the Panel's thinking about the next stage of the Review.

March 2017

Informed by the consultation and workshop, the Panel developed five evidence-gathering workstreams, each led by a Panel member.

Workstream 1: Introduction, remit and context – Professor Pam Davis-Kean

Workstream 2: Content needs – Professor Corinna Kleinert

Workstream 3: Data linkage – Professor Ray Chambers

Workstream 4: Data harmonisation – Professor Qiang Ren

Workstream 5: Training, access and discovery – Professor Leslie Davidson

A Steering Group member was 'paired' with each Panel member to provide specific input for each workstream, including information on the UK context and key contacts.

April — August 2017

The workstreams gathered evidence through the spring and summer of 2017, with each Panel member making visits to the UK to meet a diverse range of key experts, stakeholders, study directors and key staff, and users of longitudinal studies. Between them the Panel members met with over 80 individuals, in person or by telephone, and had email correspondence with many others, collecting a mix of views on key issues. Extensive written information was provided specifically for the Panel by the ESRC-funded longitudinal studies and other ESRC investments, and a wide variety of other experts. The Panel also reviewed key published reports and online material, and attended some events. While each workstream focused on particular key issues, Panel members worked closely together to ensure the sharing of issues and emerging themes across the workstreams.

In the UK, members of the Steering Group met with policy groups in government, including the Chief Scientific Advisors group, Heads of Policy Profession and Departmental Directors of Analysis. For a full list of those who were consulted, please see **Report Section 5, Appendix 1**.

July — September 2017

Each Panel member worked independently to assess the evidence for their workstream and draft their section for the report; drafts were shared among the Panel members for comment and they held regular teleconferences for sharing ideas. A small amount of further evidence collection was carried out during this period.

October — November 2017

The individual chapters were combined into a first full report draft, excluding recommendations, by Professor Davis-Kean and shared with the ESRC office and Steering Group.

The Panel convened in London on 26th and 27th October for a two-day meeting where members discussed and refined the report and their recommendations.

Professor Davis-Kean produced a revised draft of the report with full recommendations, which was circulated around the Panel for comment and agreement then shared with the ESRC office and Steering Group.

On 27th November, Professor Davis-Kean met with the Steering Group to further refine the report and ensure the UK context was fully captured.

December 2017

The final report was delivered to the ESRC office by Professor Davis-Kean.

Annex B: Descriptions of ESRC-funded longitudinal studies and other key investments

ESRC Office

October 2016

This annex provides a brief overview of each of the longitudinal studies and other key investments funded or supported in other ways by ESRC currently or recently.

There are other major funders of major longitudinal studies in the UK, including other studies that are used by social scientists, in addition to the ESRC, notably the Medical Research Council (MRC) and Wellcome (see, for example, the MRC Cohort Directory[1]). Many longitudinal studies are supported by more than one funder, sometimes for specific components, while often there are one or two core or principal funders. Included below are Understanding Society, the Centre for Longitudinal Studies (CLS) cohorts, and CLOSER, of which ESRC is the major core funder, plus the other main studies supported by the ESRC in collaboration currently or recently. A comprehensive account of the funding of each study listed is not attempted here.

This annex provides summary factual information on what each study is, what it collects and key features, as well as web links to further information. Evaluative information can be found within the discussions in the main body of the report.

Understanding Society, The UK Household Longitudinal Study

PI: Michaela Benzeval, University of Essex
Start date: 2009

Understanding Society is a household panel study that collects information annually. Through a representative sample of 40,000 households of the UK population the study shows how individuals and households are changing. By tracking changes across multiple domains such as family, employment, income, health and behaviours, Understanding Society can be used to explore short and long term dynamics in individuals' and families' lives. It builds upon and incorporates The British Household Panel Survey (BHPS) (1991-2008), the UK's first UK longitudinal household panel study.

The study collects data on an annual basis from all individuals age 16 and older from a household. In waves 1 through 4, the majority of the data was collected using face-to-face interviews. Subsequent waves implement multi-mode data collection and include bio-medical data collected from 20,000 adult study participants at waves 2 and 3.

Key features include:

- A core set of measures on income, education, households, housing, employment, behaviours, and health that are collected annually with certain content rotating across years in order to collect data on time-specific events such as the London Olympics and Brexit;
- An Innovation Panel: for experimental and methodological assessment of data collection methods and instruments. Results from experiments with survey procedures, questionnaire design, and non-experimental studies are used to advance methods of survey data collection and inform data collection strategy;
- Oversampling and Immigrant and Ethnic Minority Boost Sample: to provide a picture of ethnicity and immigration in the UK;
- Data linkage: such as linking administrative health and education records.
- Part of a family of international household panel studies which harmonises data across studies on key variables and facilitates international comparisons;
- Data deposited at UKDS and METADAC;

[1] https://www.mrc.ac.uk/research/facilities-and-resources-for-researchers/cohort-directory/

- Part of the CLOSER consortium.

Centre for Longitudinal Studies (CLS)

Centre PI: Alissa Goodman, Institute of Education, University College London

The Centre for Longitudinal Studies is responsible for four longitudinal cohort studies:

- 1958 National Child Development Study (NCDS)
- 1970 British Cohort Study (BCS70)
- Next Steps (previously Longitudinal Study of Young People in England – LSYPE)
- Millennium Cohort Study (MCS)

CLS collects the data for these studies through various methods, cleans and documents the data and deposits them with the UK Data Service or METADAC. It supports researchers to use the data and supports and promotes the sharing of best practice in survey and statistical methods. The four studies are part of the CLOSER consortium.

1958 National Child Development Study (NCDS)

PI: Alissa Goodman
Start date: 1958

The National Child Development Study follows the lives of over 17,000 people born in England, Scotland and Wales in a single week of 1958. Also known as the 1958 Birth Cohort Study, it collects information on physical and educational development, economic circumstances, employment, family life, health behaviour, wellbeing, social participation and attitudes. The age 45 sweep (in 2003) was a nurse-led bio-medical survey which collected physical measurements and biological samples, in order to learn more about how early development, environments and lifestyles affect people's health in middle age, and about the genetic underpinnings of disease.

1970 British Cohort Study (BCS70)

PI: Alice Sullivan
Start date: 1970

The 1970 British Cohort Study follows the lives of an original sample of 17,196 people born in England, Scotland and Wales in a single week of April 1970. Over the course of cohort members' lives, the BCS70 has collected information on health, physical, educational and social development, and economic circumstances among other factors. Information has been collected over time from midwives, the parents of study members, the study members themselves, and their schools, and the children of study members at age 30 sweep only, and is also linked to mortality records. The current sweep at age 46 is a biomedical sweep, consisting of a nurse home visit to collect physical and biological data.

Next Steps (LSYPE)

PI: Lisa Calderwood
Start date: 2004

Next Steps, previously known as the Longitudinal Study of Young People in England (LSYPE), follows the lives of an original 16,000 people born between 1 September 1989 and 31 August 1990. The study began in 2004, managed and funded by the Department for Education (2004-2012), when cohort members were aged 13-14, who attended state and independent schools in England and were selected to be representative of young people in England at the time the study began. The study has collected information about education, employment, economic circumstances, family life, physical and emotional health and wellbeing, social participation and attitudes. The data has also been linked to National Pupil Database (NPD) records. At wave 8 when the cohort members were 25 years old, participants were also asked for a wide range of data linkage consents in relation to education, economic circumstances, criminal behaviour and health.

Millennium Cohort Study (MCS)

PI: Emla Fitzsimons
Start date: 2000

The Millennium Cohort Study follows the lives of around 19,000 children born in the UK in 2000/01. The study has been tracking the millennium children through their early childhood years and plans to follow them into adulthood. It collects information directly from the children, their resident parents and, in two of its sweeps, older siblings. The MCS covers diverse topics such as parenting; childcare; schooling and education; daily activities and behaviour; cognitive development; child and parent mental and physical health; employment and education; income and poverty; housing, neighbourhood and residential mobility; and social capital, ethnicity and identity.

Cohort and Longitudinal Studies Enhancement Resource (CLOSER)

PI: Alison Park, Institute of Education, University College London
Start Date: 2012

CLOSER is a consortium of eight studies from the fields of medicine and social science, the UK Data Archive, and the British Library. The purpose of the consortium is to increase the use, value, and impact of the UK's longitudinal studies. The CLOSER network brings the studies together to stimulate interdisciplinary research across the major longitudinal studies, provide shared resources for research, assist with training and development for researchers in the use of longitudinal data at all career stages and to share information and expertise in longitudinal methodology.

CLOSER works across a number of different areas in order to achieve its objectives:

- CLOSER Discovery: an online metadata search platform that enables researchers to search and appraise data from the eight studies;
- Harmonising key measures, both within and across studies, to facilitate cross-study comparisons;
- Facilitating and promoting the linkage of study data to administrative records;
- Training and Capacity building activities, including introductory training for new users, workshops in advanced techniques for experienced users, and a knowledge exchange programme for longitudinal study teams;
- Programme of impact and public affairs work to promote longitudinal evidence to policymakers and practitioners.

CLOSER consortium studies:

- Hertfordshire Cohort Study
- 1946 MRC National Survey of Health and Development
- 1958 National Child Development Study
- 1970 British Cohort Study
- Avon Longitudinal Study of Parents and Children (Children of the 90s)
- Southampton Women's Survey
- Millennium Cohort Study (Child of the New Century)
- Understanding Society: The UK Household Longitudinal Study

Longitudinal Studies of Ageing

Northern Ireland Cohort for the Longitudinal Study of Ageing (NICOLA)
PI: Frank Kee, Queen's University Belfast
Start date: 2013

NICOLA is Northern Ireland's long term study of ageing in the over 50s. The study looks at health, lifestyles and financial situations of 8,500 people as they grow older, monitoring how their circumstances change over a 10-year period. It has a special focus on intergenerational poverty, transition points in ageing and the effects of diet on the ageing process. It also includes questions with specific relevance to the Northern Ireland situation.

English Longitudinal Study of Ageing (ELSA)

PI: Andrew Steptoe, UCL and a member of ESRC Research Committee
Start date: March 2002

ELSA is a unique and rich resource of information on the health, social, wellbeing and economic circumstances of the English population aged 50 and older. The ELSA includes objective and subjective data relating to health and disability, biological markers of disease, economic circumstance, social participation, networks and well-being.

Healthy Ageing in Scotland (HAGIS)

PI: David Bell, University of Stirling (ESRC does not officially fund this study)
Start date: October 2016

HAGIS is a study of ageing in the over 50s and collects data on health, economic and social circumstances.

Census Longitudinal Studies

Three Census LS Research Support Units (RSUs) were established during the 2001 Census Programme, one for each regional Longitudinal Study: England and Wales, Scotland, and Northern Ireland. In 2012, the RSUs were recommissioned along with the Census & Administrative Data Longitudinal Studies Hub (CALLS-Hub) which was designed to coordinate and provide strategic oversight of the RSUs. Following a review of the Hub in 2016 it was decided that it would not be recommissioned in 2017 at the end of the grant. The three RSUs are currently being recommissioned, with the new grants commencing on 1 April 2018 for 30 months.

- Centre for Longitudinal Study Information & User Support (CeLSIUS) (England and Wales) – based in UCL and led by Nicola Shelton, provides user support to the ONS Longitudinal Study.
- Scottish Longitudinal Study Development and Support Unit (SLS-DSU) (Scotland) – based in Edinburgh and led by Chris Dibben, assists with the development of the Scottish Longitudinal Study as well as supporting access to it.
- Northern Ireland Longitudinal Study Research Support Unit (NILS-RSU) (Northern Ireland) – based in Queen's University Belfast and led by Ian Shuttleworth, supports researchers using the NILS.
- Census and Administrative data Longitudinal Studies Hub (CALLS-HUB) – based in St Andrews and led by Allan Finlay aims to co-ordinate, harmonise and promote the work of the three LS Research Support Units, with the aim of providing a more streamlined experience for users.

Avon Longitudinal Study of Parents and Children (ALSPAC)

PI: Nic Timpson, University of Bristol
Start date: 1991

ALSPAC includes a regional sample of more than 13,000 mothers living in an area of southwest of England, who were pregnant in 1991/92. It started interviewing mothers before delivery and has continued to follow them up at irregular intervals. The main focus of ALSPAC is to find out how genetic and environmental characteristics influence health and development in parents and children.

Born In Bradford (BIB)

PI: Kate Pickett (York); co-I John Wright (Bradford Institute for Health Research)
Start date: 2007

BiB is a long term study of a cohort of 13,500 children, born at Bradford Royal Infirmary between March 2007 and December 2010, whose health is being tracked from pregnancy through childhood and into adult life. The information collected from the BiB families is being used to find the causes of common childhood illnesses and to explore the mental and social development of this generation.

Life Study

PI: Carol Dezateux, University College London
Start date: 2012, End date: 2016

Life Study was an ambitious innovative study that aimed to recruit over 80,000 babies born between 2014 and 2018 – and their families - from across the UK. The study was designed to collect information about these babies over their early lives to help us understand how early life experiences shape health and wellbeing later in childhood and adult life. Life Study was discontinued in 2016.

Annex C: Policy impact case studies

ESRC Office
December 2017

Longitudinal Studies: Evidence of Influencing Policy

The following Case Studies have been grouped by the themes identified at the Longitudinal Studies Review workshop held in January 2017. The approach to the selection of case studies is outlined at the end of the report.

Key: 🟡 🔵 🟢 🟠 🔴 🟣 🔵

Main Scientific Themes	Scientific sub-themes
Long-term effects of childhood and adult experience	Education and Skills – school, FE, HE, Lifelong
	Impact of policy on individuals, groups and communities
	Psychosocial and emotional factors
Demographic shifts and mobilities	Intergenerational dis/continuities
	Changes in employment patterns, pathways and labour markets
	Social, educational and geographical mobility differences
	Social and physical factors
Health and Wellbeing	Mental Health
Equality and Inequality	Social, economic, educational, geographic and digital/technological inequalities
Biosocial research and genomics	Biosocial research
Diversity and Identity	Political values, attitudes and voting behaviour
Ageing population	Ageing, health and well-being

1. Millennium Cohort Study research finds exclusively breastfed babies 14 times less likely to die in their early months than those not breastfed at all, providing rationale for best practice standards

Breastfed children have **at least six times greater chance of survival** in the early months than non-breastfed children. An exclusively breastfed child is **14 times less likely to die** in the first six months than a non-breastfed child, and breastfeeding drastically reduces deaths from acute respiratory infection and diarrhoea.

Findings:

- Using MCS data, researchers Quigley, Kelly and Sacker (2007) found breastfeeding to be associated with lower hospitalisation rates for respiratory infections and child diarrhoea;
- Six months of exclusive breastfeeding was associated with a 53% decrease in hospital admissions for diarrhoea and a 27% decrease in respiratory tract infections (controlling for other factors).

Impacts:

The research findings have contributed to best practice and guidance on the benefits of breastfeeding:

- UNICEF Baby Friendly Initiative Standards (2013): a worldwide programme of the World Health Organization and UNICEF which has been implemented in 134 countries. Findings have been used as part of the evidence and rationale for these standards.
- The research has been widely cited by other health organisations, such as breastfeeding information packs for UK children's centres issued in 2009 by the National Childbirth Trust, and by the British Dietetic Association in its 2013 policy statement on 'Complementary Feeding: Introduction of solid food to an Infant's Diet'.

Further information / underpinning research

- Unicef Nutrition web pages – information on breastfeeding
- REF2014. Millennium Cohort Study: building a picture of a new generation
- Dep. of Health (2009) Commissioning local breastfeeding support services. London.
- Dex, S. & Joshi, H. (2004) Millennium Cohort Study First Survey: A User's Guide to Initial Findings. London: CLS.
- British Dietetic Association (2013) policy statement, 'Complementary Feeding: Introduction of solid food to an Infant's Diet'
- NCT, 2009, Breastfeeding Pack for Children's Centres
- UNICEF: Baby Friendly Initiative: The Evidence and Rationale for the UNICEF Baby Friendly Initiative Standards (2013)

2. Research using longitudinal studies data underpins Department for Work and Pensions plans to tackle the impacts of worklessness, supporting 1.8 million children

In 2014-15 there were 1.8 million children in workless families across the UK, and in over eight out of ten cases the child was in a long-term workless family. The Department for Work and Pensions (DWP) used **Understanding Society**, a longitudinal household survey, and the **Millennium Cohort Study**, a birth cohort study, to build an understanding of the multiple disadvantages that workless families often face, and the impacts these disadvantages have on children and young people.

Findings:

- Having a parent out of work, alongside a range of associated disadvantages, has a detrimental effect on the whole family and is likely to lead to an intergenerational cycle of disadvantage.
- 75 per cent of children in workless families failed to reach the expected level at GCSE, compared to 52 per cent in lower-income working families.
- Children growing up in workless families are almost twice as likely as children in working families to fail at all stages of their education.

Approximately £47.5 billion of the Government welfare budget is spent on family benefits, income support, tax credits and the unemployed (ONS figures, 2016) while estimates made by Coles *et al* in 2010 on the life-time public finance cost of 16-18 year olds who are Not in Education, Employment or Training (NEET) were between £12-35.5 billion. Other estimates in a report commissioned by Save the Children suggest the UK GDP in 2013 would have been around £20 billion higher had action been taken to close the achievement gap between the poorest pupils and others at age 11.

Impacts:

Based on ESRC-supported research and using both Millennium Cohort and Understanding Society data, the Department for Work and Pensions (DWP) has launched a major policy initiative (presently totalling £42 million) aimed at supporting parents/carers and families who experience worklessness and economic disadvantage, with the objective of improving educational attainment and mental health outcomes for children and young people. This policy announcement uniquely recognises young people's educational attainment and mental health as primary pillars of future employability, and the importance of the longitudinal studies employed to support these conclusions and policy investment recommendations.

As an example of longitudinal study data use, the DWP describe: "We joined data on how pupils perform in key tests and exams to the **Understanding Society data** – and this has shown us for the first time what a difference it makes to children's educational attainment if they live in a workless family." This innovative, research-led policy investment is directing support to front-line professionals aimed at improving educational and mental health outcomes for children whose parents/carers experience worklessness.

Proposals based on the findings include:

- Redefining the Troubled Families Programme "to encourage a greater emphasis on tackling worklessness and issues associated with it".
- Strengthening support to help reduce relationship distress between parents/carers, whether together or separated, announcing an innovative new programme, backed initially by £30 million (April 2017), with an additional £12 million added in the November 2017 Budget Statement.

Future opportunities for further impact:

- Linking Troubled Families recommendations with the Industrial Strategy in order to support and encourage the younger generations to gain employment.
- Follow the Troubled Families Programme reviews and promote these research findings to charities working with disadvantaged families or poverty-focused e.g. Child Poverty Action Group who work to understand what causes poverty, the impact it has on children's lives, and how it can be solved – for good. Child poverty costs broader society an estimated £29 billion a year.
- Follow up with the devolved administrations and ask how they utilise the data collected on poverty. Also, local authorities may be interesting to follow in terms of how they implement the schemes and recommendations.

Further information/underpinning research

- DWP, *Improving lives: Helping Workless Families*, April 2017
- Damien Green, *Helping Workless Families: Written statement*, April 2017
- Dame Carole Black, *Drug and alcohol addiction, and obesity: effects on employment outcomes*

3. Use of the ESRC Longitudinal Studies' genetic samples and data

The ESRC funded longitudinal studies provide biomedical and genetic information in addition to social science data. NCDS, the 1958 birth cohort, is in the top four of all datasets ever used in GWAS (Genome-Wide Association Studies) discoveries, and has been used in double the number of papers of, for instance, the Wellcome Trust Case Control Consortium that is specifically designed for this type of research.

Mills and Rahal (under review) engaged in a detailed analysis of the 3,293 genetic discoveries GWAS from 2005 to 2017. GWAS form the basis of fundamental research in biology, molecular genetics, neurobiology and medical sciences.

Findings:

- In a systematic empirical overview, Mills and Rahal identified the datasets used, participant characteristics, traits under study and the funders of each type of research, by linking the databases of the GWAS Catalogue, information on publications and information on diseases.

Impacts:

- These GWAS studies have discovered key genetic variants, genes and biological pathways that play a role in specific diseases and disorders.
- Understanding the biology of diseases has in turn translated into new therapeutics and drug targets in the move towards personalised or precision medicine.
- The top datasets used were generally longitudinal cohort data that asked a wide variety of questions and were not specifically hypothesis driven, suggesting that a broader inclusion of questions incites wider usage by multiple disciplines over time.

"The usage and impact of core ESRC-funded data and birth cohorts that contains biomedical and genetic information outside of the social sciences has been remarkable. Considering that the 1958 Birth Cohort has been one the most used datasets in GWAS discoveries to date, it has made a considerable impact on international biomedical, genetic and fundamental biological research. These studies have led to fundamentally new breakthroughs in cancer (notably breast, colorectal and prostate cancer), immunity, protein measurement and the nervous system."

- Melinda Mills, Department of Sociology, University of Oxford

Further information:

- Mills, M.C. & C. Rahal (under review). The Anatomy of GWAS.
- A Genome-wide Association Study (GWAS) is a hypothesis-free data mining method using whole-genome data that tests to see whether there is a correlation between specific genetic loci and a phenotype or trait (e.g., schizophrenia, height, breast cancer).
- Access to the biomedical and genetic samples and data from the ESRC-funded longitudinal studies is managed by a specialised data access committee METADAC, rather than the UK Data Service.

4. MCS data influential in Welsh Government's commitment to tackling obesity

The proportion of the Welsh adult population which is overweight or obese currently stands at 58% and illnesses associated with obesity are estimated to cost the Welsh NHS more than £73m a year.

Findings:

- The Welsh Government used research from the **Millennium Cohort Study** as part of its evidence base to demonstrate that the proportion of adults and children who are not maintaining a healthy body weight is increasing.
- The MCS data showed that 22% of Welsh children aged three were overweight and just over 5% were obese

Impacts:

The research findings were used in The Welsh Government's All Wales Obesity Pathway in 2010 to help people achieve a healthy weight. The Pathway sets out a four-phase approach to manage and treat obesity in Wales which includes community-based prevention and early intervention services, specialist weight management services and bariatric surgery and is a tool to be used by Local Health Boards in Wales to review local policies, services and cross-departmental multi-agency working.

Future opportunities for further impact:

All Local Health Boards have prioritised obesity and continue to work with key partner organisations (public, voluntary and private) to drive this agenda. Implementation of the pathway has been driven by Local Health Boards, and delivery across Wales is being reviewed to assess current implementation. As such there have not been any formal annual updates or reviews published. Work is continuing with Local Health Boards to scope delivery and barriers.

Review of the Pathway will support the continued focus to develop this approach to tackle obesity. The Public Health (Wales) Act 2017 reinforces the Welsh Government's commitment to tackle obesity by producing a national strategy. Initial work on scoping the strategy has already begun, with a development board held on 23 October 2017, chaired by the Chief Medical Officer.

Findings could also be promoted to charities such as Weight Concern who carry out research into the causes, prevention and treatment of obesity and Obesity.org who work to better understand, prevent and treat obesity to improve the lives of those affected, through research, education and advocacy.

Further information / underpinning research

- All Wales Obesity Pathway
- http://impact.ref.ac.uk/casestudies2/refservice.svc/GetCaseStudyPDF/44326
- Speech by Health Minister Mark Drakeford
- Hansen, K. & Joshi, H. (eds). (2008) Millennium Cohort Study, Second Survey: A User's Guide to Initial Findings. London: CLS.

5. Understanding Society data essential to the DWP Implementation of Automatic Enrolment in Workplace Pensions, leading to increased total annual savings by £7.1 billion since 2012

Over the last 15 years, the Department for Work and Pensions (DWP) has developed and introduced the New State Pension and in doing so, has implemented Automatic Enrolment into workplace pensions. This has been achieved through the use of a dynamic micro-simulation model known as PenSim2, which in turn depends heavily on data from the British Household Panel Survey (now part of Understanding Society), specifically for the modules on partnership, fertility, labour market status, earnings and savings. By using large-scale datasets containing representative samples of individuals and households (either from administrative or household survey data) samples are 'grown' through time by simulating the relevant life events for each individual and each family.

Impacts:

- By 2020 over 10 million people are expected to be newly saving or saving more as a result of automatic enrolment.
- Since being initiated in 2012, more than 6.87 million workers have been automatically enrolled by 293,868 employers.
- Data collected up to April 2015 suggests that the number of eligible employees participating in a workplace pension increased to 15.1 million (75 per cent) up from 10.7 million (55 per cent) in 2012.
- The annual total amount saved by eligible employees across both sectors stands at £81.8 billion in 2015 which is an increase of £1.4 billion from 2014, and up £7.1bn since 2012.

"Without that high quality social survey data, our modelling of the impact of the reforms would have been very much weaker, and policy affecting virtually everybody in the country would have been much less well informed."

- Mike Daly, DWP Central Analysis Division

Future opportunities for further impact:

Looking to the future, it is expected that Understanding Society data will be used to update the model. Specific elements of the study that are expected to be of use to future iterations of the model (based on an audit of PenSim2 carried out in 2004) include:

- Housing Wealth
- Intergenerational Linkages
- Modelling of mortality and disability through incorporating a health module
- Pension tenures for younger individuals
- Capital Income data

Also suggested is the benefit that these improvements being made to PenSim2 would offer to other government departments including Department for Education, Department of Health and HM Revenue and Customs.

Further information / underpinning research:

- Automatic Enrolment evaluation report 2016
- Automatic Enrolment evaluation report 2015
- An assessment of PenSim2, IFS, 2004
- Workplace Pensions: Update of analysis of Automatic Enrolment, 2016

6. Longitudinal evidence supports welfare to work policy and changes common perceptions of mothers who return to work

Research using the birth cohorts (NCDS, BCS70 (and members' offspring), ALSPAC, MCS, BHPS and Children of NLSY (National Longitudinal Survey of Youth – a USA study) investigated the relationship between mother's employment and child outcomes, helping to change the prevailing presumption that children are affected from mothers going out to work, for which it found little evidence beyond the very early years, influencing the Welfare to Work policy:

"Welfare to work policy during the 2000s was very much focused on enabling mothers to enter or return to the labour market. The research, based on longitudinal studies found that, especially if complemented with access to childcare, maternal employment was not detrimental to child outcomes and was an important piece of underpinning evidence for this strategy". Jonathan Portes, Chief Economist DWP, 2002-2008

Findings:

- The research found a mother's employment, and her circumstance and characteristics to be linked prospectively to the child's outcomes at a later date.

Impacts:

The research has been influential in challenging assumptions and to changing government thinking, including research carried out by Heather Joshi in collaboration with Harriet Harman MP. This went on to support the development of policy on maternity and parental leave resulting in a report by the Smith Institute, for the government, being published 2000. The findings were cited by the Department of Trade and Industry Green Paper, Work and Parents: Competitiveness and Choice (2000, Cm 5005) in support of policies on flexible employment and leave for parents which continued to evolve into the 2010s.

The extent to which longitudinal evidence has contributed was documented by the Cabinet Office in the Independent Review on Poverty and Life Chances:

"Research has generally found small effects of early maternal employment and negative effects are insignificant if the mother goes back to work after the child is 18 months old, works part-time or flexibly and where the child is in high quality childcare during her working hours".

- Frank Field MP (2010) para 3.29

Further information/ underpinning research

- Equality in Work and Education: A series of five seminars

7. Research using National Child Development Study (NCDS) data influenced establishment of world's first universal children's savings scheme totalling £4.8 billion

Research using NCDS data carried out by Bynner and Despotidou (2000) discovered that having even very modest savings at age 23 had a wide range of beneficial economic, social and health effects 10 years later. HM Treasury used these findings to create the Child Trust Fund which will benefit approximately 6 million UK children born between 2002 and 2011 to ensure that every young person had some savings at age 18.

Findings:

- People with savings have better life chances and are happier than those without.
- Men with less than £200 (based on 1981 figures) in savings were more likely to experience unemployment than those with more savings.
- Having even very modest savings at age 23 had a very wide range of beneficial economic, social and health effects 10 years later.

Impacts:

- Findings influenced HM Treasury papers presenting options for policies designed to increase rates of saving and asset-ownership, both among lower-income households, and in generations of families in the future, such as the Saving and Assets for All: The Modernisation of Britain's Tax and Benefit System.
- It was also discussed on numerous radio programmes, where key politicians highlighted the significance of the research in underpinning thinking on asset-based savings.
- Policy-makers and think tanks from countries including the USA, France, Germany, New Zealand and Brazil have shown interest in learning from the UK's baby bonds `experience'.

Further information/ underpinning research

- Radio 4 transcript: discussion between Bynner, Blunkett and others discussing the research and Child Trust Funds.
- REF impact case study: includes references to the research.
- The Institute of Education: Case Study on the Impact of IoE Research that Underpinned the Child Trust Fund.

8. Welsh Government uses Understanding Society data to measure mental well-being among children and young people

One in four people will experience a mental health problem with the cost of mental ill health to the economy, the NHS and society as a whole at £105 billion a year (including an estimated £54 billion cost to individuals' quality of life). Left untreated, mental health conditions can result in unemployment, homelessness, the break-up of relationships and suicide. The 'Together for Mental Health' Welsh cross-government strategy launched in 2012 is a 10-year strategy which covers people of all ages, rather than through separate strategies. The most recent Delivery Plan covers the period 2016-19.

Impacts:

- Performance measures from **Understanding Society data** are being used to monitor:
    - The increased percentage of mental well-being among children and young people.
    - The mean mental well-being score for people.
- Understanding Society data will contribute to achieving the following priority areas:
    - "People in Wales are more resilient and better able to tackle poor mental well-being when it occurs".
    - "All children and young people are more resilient and better able to tackle poor mental well-being when it occurs".

Delivering the proposed actions with key stakeholders will make a positive contribution to the Welsh Government's equality objectives through a commitment to identify and meet the needs of all groups in relation to mental health. Implementation is assured through Partnership Boards at national and local levels, and progress is reported publicly through annual reports produced by the Welsh Government, and Integrated Medium Term Plans (IMTPs) of the local health boards and NHS Trusts.

Further information / underpinning research

- 'Together for Mental Health Delivery Plan' (2016-2019)
- Together for Mental Health: A strategy for Mental Health and Wellbeing in Wales
- New Investment in Mental Health Services: https://www.gov.uk/government/news/new-investment-in-mental-health-services

● 9. Age UK uses Understanding Society to create an Index of Well-being in Later Life

At a time when ageing presents one of the biggest challenges facing the UK, Age UK has created an Index of Wellbeing in Later Life which provides new and authoritative evidence about what matters for the well-being of older people. Age UK wanted to address several related issues: what was important in later life, the level of well-being amongst different groups of older people, possible reasons for low well-being and what changes in policy or practice would improve older people's lives. This Index is the first time that an overall measure of the well-being of older people has been created.

Data from Waves 1-4 of Understanding Society were used to create the Index, which is a summary measure of objective and subjective indicators grouped into domains of well-being. Age UK identified 40 indicators in five domains: personal, social, health, resources and local services.

Impacts:

Age UK is using the Index to inform their policy and practical work:

- Local Age UK branches are using early insights to think about how to target their support services at people at risk of low well-being. The Index has already been used successfully to inform a funding bid for services to offer creative and cultural activities to older people in Oxfordshire.
- The charity has engaged members of the House of Lords keen to champion evidence-based actions to target resources and effort effectively.
- There is interest from local councils who want to understand well-being among older people in their area.

"Understanding Society was chosen for the Index mainly for the number of people included in the sample, its representativeness, range of questions, UK-wide focus, and longitudinal nature. In addition, the longitudinal design of the Study means that the index can be updated to track well-being over time."

- Marcus Green, Senior Research Manager, Age UK

Further information / underpinning research

- Age UK's Wellbeing research pages

Methods and issues in identifying and developing case studies

Approach

Given the remit to identify case studies where research using data from ESRC's longitudinal investments has been ***influential in policy***, each investment and its publications were reviewed via online searches and keyword searches in specialist tools. In addition, ESRC investments themselves identified some case studies and included them in information provided for the 2017 Longitudinal Studies Review.

An initial sift identified 20 potential case studies. After office and trusted-friends review, these were further shortlisted and researched for substantial evidence of impact. As an extra review stage, they were discussed with ESRC Communications team which has significant experience in developing top-level case studies.

Challenges in developing Case Studies

The following sections show the challenges of both demonstrating policy impact and showing instrumental, conceptual and capacity impacts. These barriers are well known within the social science community.

Misunderstanding of what impact means

Many REF (Research Excellence Framework) case studies that claim instrumental impact use only the author's work as evidence. ESRC could do more to communicate a clear definition of its expectations, providing better training and more support to award-holders during their funding period. Nonetheless, strong, policy-relevant research may not influence policy, for reasons unrelated to the evidence. ESRC may need to reassess how impact is defined for data infrastructure investments in particular.

Changing political agendas

Gathering evidence during policy debate is vitally important, particularly in the case of testimonial evidence, because as time passes attribution becomes problematic and recall fades. Policy commentators can track a debate's direction, but should that direction change, retrospectively proving the link between that change of direction and the research findings can be virtually impossible. Frequent change within government and high turnover of departmental contacts are additional challenges.

Social science is not understood: social policy is contested

Impacts from social science are more about process than product. Research findings are often seen as common-sense knowledge, struggling to gain traction in policy making and recognition for achieving and demonstrating policy impacts. 'Specialist' or 'technical' research such as in medicine or environmental science do not face such problems, at least to the same extent: people don't assume their own knowledge or understanding is adequate, and instead look to an expert for evidence. Social problems can be complex and hard to track; policy debates are often politicised and ideological. Interventions in people's private social and economic lives are often seen as contentious. In comparison, debates in 'hard' science and medicine are typically less contentious, problems are more amenable to relatively simple technical solutions/fix-its, and interventions are not seen as intrusive. In addition, while most scientific, technical and medical innovations can potentially benefit everyone, much social science research is focussed on, and could benefit, disadvantaged groups. Furthermore, many medical and technical advances potentially benefit most or all of the population, while many social science findings are focused on minority groups. These barriers and challenges make getting social science research evidence heard difficult, and tracking, achieving and demonstrating policy impacts even more so.

Researchers' practice in impact tracking

Many researchers do not actively monitor indicators of impact, are unaware of how government departments use commissioned work or do not actively enquire about impact arising. Many who do try to track and evidence their own impact are attuned mainly to the national agenda, perhaps due to institutional encouragement and support for this purpose. Additionally, much social policy (broadly construed) has been decentralised to the devolved administrations, local authorities, schools etc. which makes it much more difficult to track the use of research findings.

Resource burdens

Longitudinal Studies are not funded to collect impact evidence and so requests to supply impact testimonials or track research using studies data must be limited, well-considered and proportionate. There is an opportunity to work with investments to make this process more efficient.

Information systems

Research information systems are not consistently linked to impact evidence held by funders (e.g. ResearchFish, REF impact case studies, RCUK Pathways to Impact, ORCID), nor to data on publications or outputs. How well ResearchFish is used depends on the inputter's understanding of impact. There is no box specific to the needs of explicitly capturing instrumental impacts. Guidance could be improved.

Broken links

This review found many web links consulted to be broken, making following up on policy documents and citation references in the REF case studies difficult.

Inside knowledge

'Inside knowledge' made a substantial contribution to the case studies' development; for instance, advice the ESRC office received on its association of the 'Back to Sleep' campaign with ALSPAC research allowed a serious error to be avoided.

Annex D: Summary of record linkage in CLOSER studies

CLOSER detailed record linkage status (September 2017).

Andy Boyd[1,2], Michaela Benzeval[3], Andy Wong[4], Shirley Simmons[5], Hazel Inskip[6], Emla Fitzsimons[7], Alison Park[2]

[1] ALSPAC, Population Health Sciences, Bristol Medical School, University of Bristol
[2] CLOSER, Institute of Education, University College London
[3] Understanding Society, Institute for Social and Economic Research, University of Essex
[4] MRC Unit for Lifelong Health and Ageing at UCL, University College London
[5] MRC Epidemiology Resource Centre, University of Southampton.
[6] MRC Lifecourse Epidemiology Unit, University of Southampton.
[7] Centre for Longitudinal Studies, Institute of Education, University College London

Introduction

This paper presents a detailed status update on record linkage within CLOSER studies. The tables below describe the current status of 'record linkage' within each study; distinguishing between linkages that have been achieved, those under development and those identified as being important future enhancements. The tables are formatted to report progress across the four home nations comprising the United Kingdom (which, within some domains, have different record systems or different data access requirements); although it should be noted that ALSPAC, the Hertfordshire Cohort study and the Southampton Women's Study sampled from English regions and have minimal populations in the other home nations.

Completeness

This report is based on information provided by the authors – representing all CLOSER studies – during September 2017. However mechanisms for enabling record linkage differ by study. This may lead to inconsistencies in reporting. For example, ALSPAC retain participants' residential 'easting & northing' location data within their 'Data Safe Haven' and facilitate record linkage to natural environment exposures (e.g. residential Radon exposure assessed from geological survey records). In contrast, Understanding Society make easting & northing data available via the secure laboratory functions of the UK Data Archive. In practice, this means that record linkages based on location are reported for ALSPAC, but not for Understanding Society; while in reality, activity of this type is happening at both studies.

ALSPAC				
	UK			
Area	England	Scotland	Wales	NI
Education -NPD (or equivalent)	Index children linked to NPD and data available via ALSPAC. Intend to extend to 3rd generation children.			
Children in Care, Children in Need	Index children linked to care and in need records.			
Education *HESA* *Individualised Learner Records* *Early* Years Census	One exemplar project completed and others under consideration.			
Health *Hospital records*	Pilot sample of 3,000 index children linked, application to extend this under consideration. Will consent Mothers & Fathers in 2018.			
Primary Care	Extracted records for 80% index children. Data available via UKSeRP. Will consent Mothers & Fathers in 2018.		Extracted records from some Welsh practices	

Mental Health Community Care	Application under consideration.			
Registers (birth, death, cancer)	Historically, had these on children and mothers. Application to renew under consideration.			
Other	1) Mothers linked to mammogram images. 2) Linked to midwifery and delivery records.			
Benefit and pensions **records (DWP)**	Have consented index children. Will consent Mothers & Fathers in 2018.			
Tax records **(HMRC)**	Have consented index children. Will consent Mothers & Fathers in 2018.			
Employer address	Have consented index children. Will consent Mothers & Fathers in 2018.			
Criminality	Pilot extraction of index children to Police National Computer (anonymous and cannot be linked to ALSPAC records).			
Home energy ratings **(NEED)**	No active plans.			
Vehicle registration **(DVLA)**	No active plans.			
Social media	Seeking funding for exemplar study and to investigate participant expectations.			
Voter registration	No active plans.			
Credit histories	No active plans.			
Phones/Personal Sensors	Pilot data collection underway. Planned for the future.			
GeoSpatial *IMD, Urban/Rural*	Links established for all participants.			
Census/Electoral geographies	Links established for all participants.			
Air Pollution	Linking to PMx and NOx modelled air pollution.			
Climate	No active plans.			
Green Space	Developing plans to link to records derived from satellite imagery.			
Other.	Have linked to modelled residential Radon exposure data.			

British Cohort Study (BCS70)				
	UK			
Area	England	Scotland	Wales	NI
Education -NPD (or equivalent)				
Children in Care, Children in Need				
Education *HESA* *Individualised Learner Records* *Early* Years Census				
Health *Hospital records*	Linkage in progress. Consent is collected from cohort members at Age 42 Survey 2012/13. Planned linkage. Consent is collected from cohort member's resident partner at Age 42 Survey 2012/13.	Established linkage. Consent is collected from cohort members at Age 42 Survey 2012/13. Planned linkage. Consent is collected from cohort member's resident partner at Age 42 Survey 2012/13.	Planned linkage. Consent is collected from cohort members at Age 42 Survey 2012/13. Planned linkage. Consent is collected from cohort member's resident partner at Age 42 Survey 2012/13.	
Primary Care				
Mental Health Community Care				
Registers (birth, death, cancer)	** CLS receive mortality and cause of death notifications for BCS70, but not cancer registration notifications, and intend to link mortality and cause of death notifications to BCS70 and linked HES data.*			
Other				
Benefit and pensions records (DWP)	Linkage in progress. Consent is collected from cohort members at Age 42 Survey 2012/13. Planned linkage. Consent is collected from cohort member's resident partner at Age 42 Survey 2012/13.			
Tax records (HMRC)	Linkage in progress. Consent is collected from cohort members at Age 42 Survey 2012/13. Planned linkage. Consent is collected from cohort member's resident partner at Age 42 Survey 2012/13.			
Employer address				
Criminality				
Home energy ratings (NEED)				
Vehicle registration (DVLA)				
Social media				
Voter registration				
Credit histories				
Phones/Personal Sensors				
GeoSpatial *IMD, Urban/Rural*	Various linkages to geographical identifiers, on (mainly) secure access – details available on request.			
Census/Electoral geographies				
Air Pollution				
Climate				

Green Space	
Other.	

Hertfordshire Cohort Study				
	UK			
Area	England	Scotland	Wales	NI
Education -NPD (or equivalent)	-			
Children in Care, Children in Need	-			
Education *HESA* *Individualised Learner Records* *Early* Years Census	-			
Health *Hospital records*	Extract of HES data from 1998-2010 held for all 3000 participants. Application to extend data sharing agreement is being contested by NHS Digital as consent held is no longer considered adequate			
Primary Care	Consent is in place but not currently planned			
Mental Health Community Care	-			
Registers (birth, death, cancer)	All HCS participants flagged for continuous notification of death under MR278 (a wider study of mortality in 37000 people born in Hertfordshire from 1911-1939). Section 251 cover is in place, but as with HES data, we are currently having problems with the data sharing agreement and have received no follow-up since September 2016.			
Other				
Benefit and pensions records (DWP)	-			
Tax records (HMRC)	-			
Employer address	-			
Criminality	-			
Home energy ratings (NEED)	-			
Vehicle registration (DVLA)	-			
Social media	-			
Voter registration	-			
Credit histories	-			
Phones/Personal Sensors	-			
GeoSpatial *IMD, Urban/Rural*	-			
Census/Electoral geographies	-			
Air Pollution	-			
Climate	-			
Green Space	-			
Other.	-			

Millennium Cohort Study				
	UK			
Area	England	Scotland	Wales	NI
Education -NPD (or equivalent)	Established linkage (pre-16 education) based on parental consent on behalf of cohort member at MCS4 (Age 7 Survey). *In MCS3 consent was collected for accessing cohort member school records - (England only) Planned linkage (Post-16 education). Consent will be collected from cohort members at MCS7 (Age 17 Survey).	Established linkage (pre-16 education) based on parental consent on behalf of cohort member at MCS4 (Age 7 Survey). Planned linkage (post-16 education). Consent is collected from cohort members at MCS7 (Age 17 Survey).	Established linkage (pre-16 education) based on parental consent on behalf of cohort member at MCS4 (Age 7 Survey). Planned linkage (post-16 education). Consent is collected from cohort members at MCS7 (Age 17 Survey).	Planned linkage (pre-16 education) based on parental consent on behalf of cohort member at MCS4 (Age 7 Survey). Planned linkage (post-16 education). Consent is collected from cohort members at MCS7 (Age 17 Survey).
Children in Care, Children in Need				
Education *HESA* *Individualised Learner Records* *Early* Years Census	Planned linkage. Consent is collected from cohort members at MCS7 (Age 17 Survey).			
Health *Hospital records*	Established linkage to hospital episode of delivery records. Planned linkage. Consent is collected from cohort members at MCS7 (Age 17 Survey) (to override parental consent collected at MCS4).	Established linkage to hospital episode of delivery records. Established linkage 0-14 years. Pending onward sharing agreement. Planned linkage: consent is collected from cohort members at MCS7 (Age 17 Survey).	Established linkage to hospital episode of delivery records. Established linkage 0-14 years. Pending onward sharing agreement. Planned linkage: consent is collected from cohort members at MCS7 (Age 17 Survey)	Established linkage to hospital episode of delivery records. Planned linkage. Consent is collected from cohort members at MCS7 (Age 17 Survey)
Primary Care	Planned linkage. Consent is collected from cohort members at MCS7 (Age 17 Survey)			
Mental Health *Community Care*	*Under consideration at present			

Registers (birth, death, cancer)	* MCS receive regular updates/death notification from NHS prior fieldwork and under Section 251 for the entire cohort. *Consent was collected from mothers of cohort members at MCS1 to access birth registration
Other	
Benefit and pensions records (DWP)	Planned linkage. Consent is collected from cohort members at MCS7 (Age 17 Survey). Linkage in progress. Consent is collected from cohort members' main parent and partner at MCS4 (and MCS5).
Tax records (HMRC)	Planned linkage. Consent is collected from cohort members at MCS7 (Age 17 Survey). Linkage in progress. Consent is collected from cohort members' main parent and partner at MCS4 (and MCS5).
Employer address	
Criminality	Planned linkage. Consent is collected from cohort members at MCS7 (Age 17 Survey).
Home energy ratings (NEED)	
Vehicle registration (DVLA)	
Social media	
Voter registration	
Credit histories	
Phones/Personal Sensors	
GeoSpatial *IMD, Urban/Rural*	Various linkages to geographical identifiers, on (mainly) secure access – details available on request.
Census/Electoral geographies	
Air Pollution	Ongoing work on linkages to MEDix air pollution deciles.
Climate	
Green Space	Existing linkage to green space deciles (sweeps 1-5) at LSOA (England) and ward (UK) levels.
Other.	

MRC National Survey of Health and Development				
	UK			
Area	England	Scotland	Wales	NI
Education -NPD (or equivalent)	No active plans.			
Children in Care, Children in Need	No active plans.			
Education *HESA* *Individualised Learner Records* *Early* Years Census	No active plans.			
Health *Hospital records*	Application approved.	Application under consideration.		
Primary Care	Plans in development.			
Mental Health *Community Care*	Planned for the future.			
Registers (birth, death, cancer)	Historically, had these on all participants. Application approved.	Historically, had these on all participants. Application under consideration.		
Other	Women linked to mammogram images.			
Benefit and pensions records (DWP)	No active plans.			
Tax records (HMRC)	No active plans.			
Employer address	No active plans.			
Criminality	No active plans.			
Home energy ratings (NEED)	No active plans.			
Vehicle registration (DVLA)	No active plans.			
Social media	No active plans.			
Voter registration	No active plans.			
Credit histories	No active plans.			
Phones/Personal Sensors	Planned for the future.			
GeoSpatial *IMD, Urban/Rural*	Links established for all participants with valid postcodes.			
Census/Electoral geographies	Links established for all participants with valid postcodes.			
Air Pollution	Linking to PMx and NOx modelled air pollution, with historical studies linked to ‘black smoke’ and SO.			
Climate	No active plans.			
Green Space	Planned for the future.			
Other.				

National Child Development Study (NCDS)				
	UK			
Area	England	Scotland	Wales	NI
Education -NPD (or equivalent)				
Children in Care, Children in Need				
Education *HESA* *Individualised Learner Records* *Early* Years Census				
Health *Hospital records*	Linkage in progress. Consent is collected from cohort members at Age 50 Survey 2008/2009. Planned linkage. Consent is collected from cohort members' resident partner at Age 50 Survey 2008/2009.	Established linkage. Consent is collected from cohort members at Age 50 Survey 2008/2009. Planned linkage. Consent is collected from cohort members' resident partner at Age 50 Survey 2008/2009.	Planned linkage. Consent is collected from cohort members at Age 50 Survey 2008/2009. Planned linkage. Consent is collected from cohort members at Age 50 Survey 2008/2009.	
Primary Care				
Mental Health *Community Care*				
Registers (birth, death, cancer)	** CLS receive mortality and cause of death notifications for NCDS, but not cancer registration notifications, and intend to link mortality and cause of death notifications to NCDS and linked HES data.*			
Other				
Benefit and pensions records (DWP)	Linkage in progress. Consent is collected from cohort members at Age 50 Survey 2008/2009. Planned linkage. Consent is collected from cohort member's resident partner at Age 50 Survey 2008/2009.			
Tax records (HMRC)	Linkage in progress. Consent is collected from cohort members at Age 50 Survey 2008/2009. Planned linkage. Consent is collected from cohort member's resident partner at Age 50 Survey 2008/2009.			
Employer address				
Criminality				
Home energy ratings (NEED)				
Vehicle registration (DVLA)				
Social media				
Voter registration				
Credit histories				
Phones/Personal Sensors				

GeoSpatial *IMD, Urban/Rural*	Various linkages to geographical identifiers, on (mainly) secure access – details available on request
Census/Electoral geographies	
Air Pollution	
Climate	
Green Space	
Other.	

Southampton Women's Study				
	UK			
Area	England	Scotland	Wales	NI
Education -NPD (or equivalent)				
Children in Care, Children in Need				
Education *HESA* *Individualised Learner Records* *Early* Years Census	Consent to obtain education data from the local authority but not obtained yet as waiting for GCSE results			
Health *Hospital records*	Plans to obtain consent at age 17 to cover adult linkage of the offspring to health records			
Primary Care	As above			
Mental Health Community Care	Possibly as above			
Registers (birth, death, cancer)	Possibly as above			
Other	Consent for child health to provide contact details and this has been used on occasions			
Benefit and pensions records (DWP)				
Tax records (HMRC)				
Employer address				
Criminality				
Home energy ratings (NEED)				
Vehicle registration (DVLA)				
Social media				
Voter registration				
Credit histories				
Phones/Personal Sensors				
GeoSpatial *IMD, Urban/Rural*	IMD data obtained for different subgroups of the study			
Census/Electoral geographies				
Air Pollution				
Climate				
Green Space				
Other.				

Understanding Society				
	UK			
Area	England	Scotland	Wales	NI
Education -NPD (or equivalent)	Data available for users in UKDS Secure lab from w1 consents; w4 consent	Consents collected, Linkage and sharing contract signed, (for UKDS Secure lab), 3rd part matching process being prepared	Consents collected, Agreement being negotiated	Not started
Education *HESA* *Individualised Learner Records* *Early* Years Census	Consents in field	Consents in field	Consents in field	
Health *Hospital records*	Consents collected, Stalled linkage application	Consents collected, Linkage and sharing contract signed (for UKDS Secure lab), 3rd part matching process being prepared	Consents collected, Agreement being negotiated	
Benefit and pensions records (DWP)	Consents collected, agreement with DWP to be signed off			
Home energy ratings (NEED)	Consents in field			
Tax records (HMRC)	Consents collected, agreement negotiation in queue behind CLOSER project			
Vehicle registration (DVLA)	Consents collected, agreement with DfT to be signed off			
Social media	Consents in field, IP only			NI not in IP
Employer address	Consents in field, IP only			
Voter registration	Consents in field, IP only			
Credit histories	Consents collected in IP; agreement waiting to be signed			

Annex E: Summary of data harmonisation in CLOSER studies

R French, CLOSER, October 2017

https://www.closer.ac.uk/about/areas-work/data-harmonisation/ (October 2017)

Existing harmonisation

Work Package #	Title	Studies involved		Topics/measures	Deposit	Stage
WP1	Harmonisation of measures of body size and body composition	1946 NSHD 1958 NCDS 1970 BCS	ALSPAC MCS	Obesity BMI Height / Weight	https://discover.ukdataservice.ac.uk/series/?sn=2000111	Completed
WP2	Harmonisation of socio-economic status and qualifications	1946 NSHD 1958 NCDS 1970 BCS	ALSPAC MCS	Childhood and Adulthood: Education Social Class Income	Awaiting deposit	Completed
WP4	Harmonising measures of senses and behaviours	1946 NSHD 1958 NCDS 1970 BCS ALSPAC		Visual Health Social Inequalities	Awaiting deposit	Completed
WP9	Prospective associations between childhood environment and adult mental wellbeing	1946 NSHD 1958 NCDS 1970 BCS		Child rearing Parental interest in education Parental divorce Parental health Parent and child relationships	Awaiting deposit	Completed
WP13	Overcrowding and health: Methodological innovation for socioeconomic measures in longitudinal studies	1958 NCDS 1970 BCS Understanding Society		Overcrowding Income Tenure Good health Life satisfaction Persons per room Bedroom standard	Awaiting deposit	Completed

Future harmonisation - Extension work packages (2018-19)

Work Package #	Title	Studies involved		Topic	Completion date
WP15	Socioeconomic differentials in physical activity by age and cohort: enhancing the CLOSER cohort resource to inform research, policy and practice	1946 NSHD 1958 NCDS 1970 BCS	ALSPAC MCS Understanding Society	Physical activity Sedentary behaviour Socio-economic position	April 2019
WP16	Maximising the take up of mental health measures from UK cohorts and longitudinal studies	HCS 1946 NSHD 1958 NCDS 1970 BCS SWS ALSPAC MCS Understanding Society	TEDS ELSA Born in Bradford Whitehall II UK Biobank E-Risk Longitudinal Twin Study	Mental health and wellbeing measures	April 2019
WP17	Scoping existing dietary data available in CLOSER to support cross-cohort research questions	HCS 1946 NSHD 1958 NCDS 1970 BCS SWS	ALSPAC MCS Understanding Society TILDA	Dietary intake information e.g. food frequency questionnaires diet diaries	April 2019
WP18	The creation of a life course methylome through data harmonization in CLOSER studies	1958 NCDS 1970 BCS SABRE Lothian Birth Cohort TwinsUK		DNA methylation samples	April 2019
WP19	Assessment and harmonisation of cognitive measures in British birth cohorts	1946 NSHD 1958 NCDS 1970 BCS ALSPAC MCS		Cognitive Measures, e.g. Reading Tests Reading comprehension Vocabulary Number Skills	April 2019

Annex F: The Representativeness of the CLS cohorts and Understanding Society

ESRC Office, December 2018

This Annex contains information on the representativeness of the cohorts held at the Centre for Longitudinal Studies (CLS) and the household panel study, Understanding Society. The CLS cohorts include:

- National Child Development Study (NCDS) – 1958 birth cohort
- British Cohort Study (BCS70) – 1970 birth cohort
- Next Steps (previously known as Longitudinal Study of Young People in England)
- Millennium Cohort Study (MCS) – 2000 birth cohort

A selection of characteristics of participants in each study have been investigated and compared against the UK population.

Annex Contents

Information on Methods

The population comparison figures used in this annex reflect the coverage of the study sample where possible. For example, in some of tables the population comparison figures for NCDS cover people in the population aged 55-59 living in England, Wales and Scotland (not Northern Ireland). This means these comparison figures broadly reflect the sample of the population the NCDS is able and intended to represent as a birth cohort of individuals born in 1958 in England, Wales and Scotland.[2]

Demographic figures from each study were based on the most recent sweep of data collection.[3] Details of each study's sample and sweep are shown in Table 1; population comparisons have been matched as closely to these characteristics

[2] A point to note is that the population figures contain all non GB-born immigrants while the NCDS samples, by design, include only non-GB born immigrants who were traced through schools and joined the survey during the childhood sweeps (up to age 16). However in other tables it has been possible to include population-wide, rather than age-specific population demographics, and in these the comparability between the birth cohort samples and these population figures is much reduced.

[3] It is important to note that figures from CLS for NCDS and BCS70 are not adjusted for attrition/other forms of missing data, whereas figures from Next Steps and MCS are (in some tables) adjusted ("weighted") for the complex design of the surveys as well as attrition/unit non response. Simple corrections – through weighting or multiple imputation are also possible for NCDS and BCS70, but were not requested for the purposes of this Annex, which have been shown to align these characteristics to their relevant population totals. The CLS 'missing data strategy' describes these corrections and how they restore their samples to representativeness. See

as possible with available population statistics. Where possible, all comparison figures have been sourced from freely available datasets.

All figures have been shown as a percentage of the total 'target' population (e.g. England, Scotland and Wales for NCDS and only England for Next Steps). Percentages are to 1 decimal point; they may not necessarily add to 100%.

This is a complicated task. The figures are accurate as far as has been possible in the confines of this exercise. In some cases the categories used differ in different studies, for example ethnic group, and the condensed version presented here does not provide granular information on some groups of interest who may be under represented.

Table 1: Basic details of study sweeps

Study	Understanding Society	NCDS	BCS70	Next Steps	MCS
Participant year of birth	n/a	1958	1970	1989-90	2000
Year of last sweep	2014–2016 (Wave 6)	2013–2014	2012–2013	2015–2016	2015
Participant age at last sweep	All ages covered	55	42	25	14
Countries covered	England Scotland Wales N. Ireland	England Scotland Wales	England Scotland Wales N. Ireland	England	England Scotland Wales N. Ireland

Data collected from CLS and Understanding Society

Data on selected demographics at the last available sweep of all CLS studies and Understanding Society were requested and were received in September 2017.

Figures from CLS for NCDS and BCS70 are not adjusted for attrition/other forms of missing data, whereas figures from Next Steps and MCS are adjusted ("weighted") for the complex design of the surveys as well as attrition/unit non response.

Where provided, notes about the data have been included below each table.

Country of residence

Understanding Society

Persons (all ages)	UK Population % Mid-2015	UKHLS Wave 6	
		Non Wtd %	Wtd %
England	84.1	77.3	84.2
Wales	4.8	7.3	4.9
Scotland	8.3	8.5	8.1
Northern Ireland	2.8	6.9	2.8

Understanding Society figures based on all persons in participating households of wave 6 (household grid and household questionnaire completed) in the General Population Sample, Ethnic Minority Boost Sample and BHPS sample (including Scotland, Wales and Northern Ireland boost samples) combined. Weighted proportions use the weight `f_psnenub_xw`.

George B. Ploubidis (2017) Presentation on CLS 'missing data strategy', at the Royal Statistical Society, 26th October 2017 , event on Large longitudinal studies: design and methodology.

Population figures are from ONS Population Estimates for UK, England and Wales, Scotland and Northern Ireland (Mid-2015, MYE1, all countries); estimates cover the population living in communal establishments as well as in households.

Centre for Longitudinal Studies

Country	NCDS		BCS70		Next Steps		MCS	
	Pop %	Study %	Pop %	Study %	Pop %	Study %	Pop %	Study %
England	85.3	78.3	84.3	80.2	84.5	100	84.5	63.5
Wales	5.2	10.7	4.5	4.8	4.5	0	4.6	14.3
Scotland	9.5	4.9	8.3	9.3	8.3	0	7.7	12.1
Northern Ireland			2.8	3.5	2.7	0	3.2	10.0
Born in GB but country not known	-	1.8	-	-	-	-	-	-
Born outside Great Britain (Post-birth immigrant)		4.3		2.1		0		0

CLS figures show percentage of sample born in each constituent country of the UK and those born outside of Great Britain.

Population estimates show the percentage of the population resident in each constituent country. Next Steps comparisons do not match the sample geographically (i.e. comparison figures are not restricted to only England) in order to compare the spread of the population of persons aged 25-29 with the Next Steps Sample.

Population estimates data sources and details:
NCDS *= ONS Population Estimates for UK, England and Wales, Scotland and Northern Ireland; Mid-2014, MYE1 (Population Summary); persons aged 55-59, excluding Northern Ireland*
BCS70 *= ONS Population Estimates for UK, England and Wales, Scotland and Northern Ireland*; Mid-2013, *MYE1 (Population Summary); persons aged 40-44, all countries*
Next Steps =ONS *Population Estimates for UK, England and Wales, Scotland and Northern Ireland; Mid-2016, MYE1 (Population Summary); persons aged 25-29, all countries*
MCS = ONS *Population Estimates for UK, England and Wales, Scotland and Northern Ireland; Mid-2015, MYE1 (Population Summary); persons aged 10-14, all countries*

Age

Understanding Society

Persons (all ages)	UK Population % (Mid- 2015)	UKHLS Wave 6	
		Non Wtd %	Wtd %
0-4	6.2	5.7	5.5
5-9	6.1	6.7	6.0
10-14	5.5	6.7	5.8
15-19	5.9	6.7	5.9
20-24	6.6	6.3	6.1
25-29	6.8	5.3	5.6
30-34	6.7	5.7	5.9
35-39	6.3	5.9	5.7
40-44	6.6	6.9	6.5
45-49	7.1	7.5	7.4
50-54	7.0	7.1	7.3
55-59	6.1	6.4	6.6
60-64	5.4	5.8	6.1
65-69	5.6	5.9	6.4
70-74	4.2	4.4	4.6
75-79	3.3	3.3	3.6
80-84	2.4	2.2	2.7
85-89	1.5	1.1	1.6
90 +	0.9	0.5	0.8

Understanding Society distributions are based on all persons in participating households of wave 6 (household grid and household questionnaire completed) in the General Population Sample, Ethnic Minority Boost Sample and BHPS sample (including Scotland, Wales and Northern Ireland boost samples) combined. Weighted proportions use the weight `f_psnenub_xw`*.*

Population figures are from ONS Population Estimates for UK, England and Wales, Scotland and Northern Ireland (Mid-2015, MYE1, UK population by age group).

Sex

Understanding Society

Sex	UK Population % (Mid-2015)	Understanding Society	
		Non Wtd %	Wtd %
Male	49.3	48.3	48.9
Female	50.7	51.7	51.5

Population figures are from ONS Population Estimates for UK, England and Wales, Scotland and Northern Ireland (Mid-2015, MYE1, UK population, Males vs Females).

Centre for Longitudinal Studies

	NCDS		BCS70		Next Steps			MCS		
	Pop %	Study %	Pop %	Study %	Pop %	Study %		Pop %	Study %	
						Non Wtd	Wtd		Non Wtd	Wtd
Male	49.4	48.5	49.2	48.0	51.1	44.4	50.8	51.1	50.2	52.4
Female	50.6	51.5	50.8	52.0	48.9	55.6	49.2	48.9	49.8	47.6

Population estimates sources:
NCDS *= ONS Population Estimates for UK, England and Wales, Scotland and Northern Ireland; Mid-2014, MYE2, number of males and females aged 55 in England, Wales and Scotland (not including Northern Ireland)*
BCS70 *= ONS Population Estimates for UK, England and Wales, Scotland and Northern Ireland; Mid-2013, MYE2, number of males and females aged 42 in UK (England, Wales, Scotland and Northern Ireland)*
Next Steps *= ONS Population Estimates for UK, England and Wales, Scotland and Northern Ireland; Mid-2016, MYE2, number of males and females aged 25 in England only*
MCS *= ONS Population Estimates for UK, England and Wales, Scotland and Northern Ireland; Mid-2015, MYE2, number of males and females aged 14 in the UK (England, Wales, Scotland and Northern Ireland)*

Ethnic Group

Understanding Society

Ethnic Group	UK Population %	Ethnic Group	UKHLS Wave 6	
			Non Wtd %	Wtd %
White	88.5	White	79.7	91.5
Indian	2.4	Indian	4.3	1.8
Pakistani/Bangladeshi	2.2	Pakistani	3.8	1.1
		Bangladeshi	2.1	0.5
		Other Asian	1.8	1.2
Black	2.8	Black	5.4	2.1
Mixed	0.9	Mixed	1.9	1.1
Other	3.1	Other	1.1	0.7
No info	0.1			

Understanding Society figures are based on all persons aged 16 or over in wave 6 completing the individual interview or having a proxy interview completed on their behalf. Weighted proportions use the weight `f_ indpxui_xw`*.*

Population estimates source: ONS Annual Population Survey (accessed via Nomis); Geography: UK; Date: January – December 2015; Table T18 Ethnicity by age (cells: T18:1 T18:4, T18:7, T18:10, T18:13, T18:16, T18:19); All ages

Centre for Longitudinal Studies

	NCDS		BCS70		Next Steps		MCS	
	Pop %	Study %	Pop %	Study %	Pop %	Study %	Pop %	Study %
White	94.1	97.8	85.0	95.3	81.6	68.6	82.4	76.5
Indian	1.6	0.4	3.2	1.0	3.9	6.5	2.2	2.6
Pakistani/Bangladeshi	0.8	0.3	3.3	0.7	4.0	10.7	4.2	7.1
Black	1.6	0.6	3.2	0.7	3.9	7.3	4.6	3.1
Mixed	0.3	0.3	1.1	0.4	1.4	4.7	2.8	4.6
Other	1.5	0.7	4.2	0.9	5.1	2.2	3.6	2.4
Insufficient info	0.1	-	<0.1	1.0	0.1	0.1	0.1	3.8

Population statistics sources:
NCDS *= ONS Annual Population Survey (accessed via Nomis); Geography: England, Scotland, Wales; Date: January – December 2014; Table T18 'Ethnicity by age' (cells: T18:85, T18:88, T18:91, T18:94, T18:97, T18:100, T18:103); Aged 50+.*
BCS70 *= ONS Annual Population Survey (accessed via Nomis); Geography: UK; Date: January – December 2013; Table T18 'Ethnicity by age' (cells: T18:64, T18:67, T18:70, T18:73, T18:76, T18:79, T18:82); Aged 25-49.*
Next Steps *= ONS Annual Population Survey (accessed via Nomis); Geography: England; Date: January – December 2016; Table T18 'Ethnicity by age' (cells: T18:64, T18:67, T18:70, T18:73, T18:76, T18:79, T18:82); Aged 25-49.*
MCS *= ONS Annual Population Survey (accessed via Nomis); Geography: UK; Date: January – December 2015; Table T18 'Ethnicity by age' (cells: T18:22, T18:25, T18:28, T18:31, T18:34, T18:37, T18:40); Aged 16-19.*

CLS ethnic group categories (Some categories provided by CLS have been combined to match the Annual Population Survey categories):
NCDS*= 'Other' includes Chinese*
BCS70 *= 'White' includes British, Irish and Other White; 'Black' includes Caribbean, African, Other Black; 'Mixed' includes White & Black Caribbean, White & Black African, White & Asian and Other Mixed Race; 'Other' includes Other Asian, Indian/Pakistani, Chinese and Other Ethnic Group*
Next Steps *= 'Black' includes Black Caribbean and Black African; 'Other' includes Chinese*

MCS = 'Black' includes Black Caribbean, Black African and Other Black; 'Other' includes Other Asian, Chinese and Other Ethnic Group

Immigrant Status (Country of Birth)

Understanding Society

Country of birth (all ages)	UK Population %	Understanding Society	
		Non Wtd %	Wtd %
UK	86.6	81.1	89.1
Ireland	0.6	0.6	0.5
Poland	1.3	0.9	0.8
Other EU	3.1	2.6	2.3
Other Europe	0.5	0.4	0.3
Africa	2.2	4.1	2.1
Americas	1.1	1.8	1.0
Indian	1.2	2.3	1.1
Pakistan	0.8	2.3	0.7
Bangladesh	0.3	1.3	0.3
Other Asia	1.9	1.7	1.2
Australasia	0.3	0.3	0.3
Other	<0.1	0.6	0.4

Understanding Society figures are based on all persons aged 16 or over in wave 6 completing the individual interview or having a proxy interview completed on their behalf. Weighted proportions use the weight `f_ indpxui_xw`.

Population estimates source: ONS Dataset: Population of the UK by country of birth and nationality; January to December 2015; some categories have been separated/combined to match the categories from Understanding Society.

Centre for Longitudinal Studies

	NCDS		BCS70		Next Steps		
	Pop %	Study %	Pop %	Study %	Pop %	Study %	
						Non Wtd	Wtd
Born in UK	83.1	96.2	84.0	98.1	80.3	91.0	93.7
Immigrant	16.9	3.8	16.0	1.9	19.7	6.3	4.6
Unknown	-	-	-	-	-	2.7	1.7

PLEASE NOTE: Population statistics do not match the cohort samples on age and instead cover all persons aged 16-64. They are matched on geography and year of data collected.
Data on MCS was not available from CLS; data on those aged under 16 was not readily available from APS using Nomis.

Population statistics sources:
***NCDS** = ONS Annual Population Survey – Regional – Country of Birth (accessed via Nomis); Geography: England, Scotland, Wales; Date: January-December 2014; All persons – aged 16-64.*
***BCS70** = ONS Annual Population Survey – Regional – Country of Birth (accessed via Nomis); Geography: UK; Date: January – December 2013; All persons – aged 16-64*
***Next Steps** = ONS Annual Population Survey – Regional – Country of Birth (accessed via Nomis); Geography: England; Date: January – December 2016; All persons – aged 16-64*

Persons per household

	UK Population % (2015)	Understanding Society	
		Non Wtd %	Wtd %
1 person	28.6	25.7	32.2
2 people	35.0	33.4	34.4
3 people	16.0	15.9	14.0
4 people	14.1	15.5	13.0
5 people	4.4	6.0	4.5
6 or more	2.0	3.5	1.9

Understanding Society figures are based on all participating households. Weighted proportions use the weight `f_ hhdenui_xw`.

Population estimate source: ONS Dataset: Families and households in the UK; (based on the Labour Force Survey), Table 5: Households by size; 2015.

Centre for Longitudinal Studies

	NCDS		Next Steps			MCS		
People in household	Pop %	Study %	Pop %	Study %		Pop %	Study %	
				Non Wtd	Wtd		Non Wtd	Wtd
1	28.4	12.7	28.3	**20.8**	**19.7**	28.6	**13.3**	**14.1**
2	35.0	44.1	34.9	**28.2**	**29.2**	35.0	**43.7**	**42.2**
3	16.1	24.8	16.1	**20.7**	**22.4**	16.0	**25.7**	**25.4**
4	13.9	13.8	14.1	**16.2**	**17.0**	14.1	**11.5**	**12.1**
5	4.5	3.4	4.7	**7.6**	**7.1**	4.4	**3.6**	**3.6**
6 or more	2.1	1.2	1.9	**6.5**	**4.6**	2.0	**2.2**	**2.6**

PLEASE NOTE: Population statistics do not match the cohort samples on age or geography, and instead cover all households in the UK. They are matched on year of data collected.
CLS data shows the number of people currently living in household (incl. cohort member).

Population estimate source: ONS Dataset: Families and households in the UK; (based on the Labour Force Survey), Table 5: Households by size; NCDS = 2014; Next Steps = 2016, MCS = 2015.

BCS70 – Number of children

UK Population 2013 (%)		BCS70 (%)	
No children	28.9	No children	22.1
1 dependent child	14.0	1 child	18.1
2 dependent children	11.8	2 children	39.0
3 or more dependent children	4.3	3 or more children	20.9
Non-dependent children	10.8		

BCS70 data shows the number of own children of BCS70 cohort members (in household or absent).

PLEASE NOTE: Population statistics do not match the cohort sample on age or category, and instead cover all households in the UK in 2013.

Population estimate source: ONS Dataset: Families and households in the UK; (based on the Labour Force Survey), Table 1: Number of families (extracted: No children and Non-dependent children) and Table 3: Number of families with dependent children (extracted: Numbers of dependent children); 2013.

Education

Population		Understanding Society		
Persons aged 16-64	%	Persons aged 16-64	Non Wtd %	Wtd %
Degree or equivalent	28.3	Degree	27.68	27.66
Higher education below degree	9.0	Other higher degree	11.72	11.71
A level or equivalent	22.9	A level or equivalent	24.37	25.01
GCSE grades A-C or equivalent	21.7	GCSE or equivalent	22.02	22.42
Other qualifications (GCSE)	9.2	Other qualification	7.26	7.25
No qualifications (GCSE)	8.9	No qualification	6.95	5.96

Understanding Society figures are based on all persons aged 16-64 in participating households in wave 6 (household grid and household questionnaire completed) in the General Population Sample, Ethnic Minority Boost Sample and BHPS sample (including Scotland, Wales and Northern Ireland boost samples) combined. Weighted proportions use the weight `f_ psnenub_xw`.

Population estimates source: ONS Annual Population Survey (accessed via Nomis); Geography: UK; Date: January – December 2015; Variable: Qualifications (GCSE) by age; persons aged 16-64

Centre for Longitudinal Studies

NCDS

Population		NCDS	
Persons aged 50-64	%	(Age 55)	%
NVQ 4+	33.6	NVQ 4+	37.6
NVQ 3 only	13.3	NVQ 3	17.5
NVQ 2 only	13.8	NVQ 2	24.2
NVQ 1 only	12.4	NVQ 1	10.3
Trade apprenticeships	5.8		
No qualifications (NVQ)	13.6	None	8.3
Other qualifications (NVQ)	7.4	Not enough info	2.0

BCS70

Population		BCS70	
Persons aged 40-49	%	(Age 42)	%
NVQ 4+	38.5	NVQ 4+	41.3
NVQ 3 only	13.6	NVQ 3	14.7
NVQ 2 only	16.0	NVQ 2	24.6
NVQ 1 only	13.9	NVQ 1	7.6
Trade apprenticeships	3.6		
No qualifications (NVQ)	8.2	None	11.8
Other qualifications (NVQ)	6.2	Not enough info	0.1

Next Steps

Population		Next Steps		
Persons aged 25-29	%	(Age 25)	Non Wtd %	Wtd %
NVQ 4+	43.0	NVQ 4+	42.0	34.0
NVQ 3 only	13.8	NVQ 3	19.5	16.4
NVQ 2 only	14.8	NVQ 2	21.8	25.0
NVQ 1 only	12.5	NVQ 1	9.6	15.6
Trade apprenticeships	2.4	Other academic qualifications	0.1	0.1
No qualifications (NVQ)	6.6	None	7.1	9.0
Other qualifications (NVQ)	6.9			

CLS figures show educational attainment by age of last sweep from an academic or vocational qualification. MCS have not yet collected this data (the cohort is too young).

CLS categories:
NVQ 1 *= foundation GNVQ, three to four GCSEs at grades D-E, Business & Technology Education Council (BTEC) first certificate;*
NVQ 2 *= four or five GCSEs at grades A*–C, BTEC first diploma;*
NVQ 3 *= two or more A levels, BTEC Ordinary National Diploma (OND), City & Guilds Advanced Craft;*
NVQ 4 *= BTEC Higher National Certificate (HNC) or Higher National Diploma (HND), or City & Guilds Full Technological Certificate / Diploma;*
NVQ 5 *= Master's degree, Postgraduate certificate, Postgraduate diploma, Doctorate*

Population estimates sources:
NCDS *= ONS Annual Population Survey (accessed via Nomis); Geography: England, Wales, Scotland; Date: January – December 2014; Variable: Qualifications (NVQ) by age; persons aged 50-64*
BCS70 *= ONS Annual Population Survey (accessed via Nomis); Geography: UK; Date: January – December 2013; Variable: Qualifications (NVQ) by age; persons aged 40-49*
Next Steps *= ONS Annual Population Survey (accessed via Nomis); Geography: England; Date: January – December 2016; Variable: Qualifications (NVQ) by age; persons aged 25-29*

Gross weekly household income

Understanding Society (for individuals with household income ≥ £115 per week)

Income Range (per week)	Population (HBAI)		Understanding Society	
	%	Cumulative %	%	Cumulative %
£115-£250	7.42	7.42	6.3	6.3
£250-£365	11.07	18.49	10.41	16.71
£365-£500	13.18	31.67	13	29.7
£500-£615	10.35	42.02	10.16	39.87
£615-£923	22.15	64.17	22.35	62.22
£923-£1231	14.43	78.6	15.57	77.79
£1231-£1846	13.32	91.92	14.31	92.1
£1846+	8.08	100	7.9	100

Understanding Society figures are based on all participating households in wave 6.

HBAI (Households Below Average Income) is the source for official UK statistics on the income distribution. The HBAI corresponds to a financial year (April to March) and an Understanding Society wave to two calendar years. To account for differences in the fieldwork period, we pool two consecutive HBAI data sets when comparing to a single Understanding Society wave (wave 6, 2014-2015). All figures are expressed in 2014-15 prices using the bespoke monthly CPI price index used in the official UK income statistics and produced by the Office for National Statistics.

Note that comparative income figures are not provided for the CLS studies because ONS was not able to provide figures as the sample sizes are too small to provide meaningful figures.

Centre for Longitudinal Studies

Income Range (per week)	NCDS		BCS70		MCS (parents)	
	Pop.	Study (age 55)	Pop.	Study (age 42)	Pop.	Study (child age 12)
£0-£115		8.9		5.8		7.3
£115-£250		11.0		7.4		10.0
£250-£365		11.6		10.3		13.5
£365-£500		13.5		13.8		18.4
£500-£615		11.3		11.9		12.5
£615-£923		22.8		25.0		22.6
£923-£1231		10.4		12.6		8.3
£1231-£1846		6.02		7.9		5.2
£1846-£3520		3.4		3.5		1.8
£3520-£5175		0.3		0.7		0.3
£5175-£35000		0.6		0.9		0.1
£35000-£460000		0		0.2		0

Annex G: A brief note on UK legislation for data processing and sharing for research in the social sciences

ESRC Office
November 2017

Broadly speaking, there are three key legislative frameworks that influence how data about people are and will be processed and shared for academic research and research resources including longitudinal studies in the UK.

The **Data Protection Act (DPA) 1998** regulates the processing of personal data (i.e., any data that can be used to identify a living individual) and sensitive personal data, which covers all health data and other sensitive data about an individual's race, ethnicity, politics, religion, sex life and trade union status. The legislation aims to protect the rights of individuals about whom personal data are obtained, stored, processed or supplied and it is tied to EU Data Protection Law. The Act sets out 8 data protection principles but Section 33 provides certain exemptions in respect of the processing (or further processing) of personal data for "research, history and statistics".

On 25th May 2018, the UK data protection law will change in accordance with changes in EU Data Protection law, when the **EU General Data Protection Regulation** (GDPR) will apply across the EU. UK legislation that will implement the GDPR and replace the DPA 1998 is being developed, the **Data Protection Bill**, and will specify, among other things, the legal bases for processing personal data and exceptions for research purposes. However it is worth highlighting that the new law will not change the requirement on data controllers to have a lawful basis in order to process personal data but it will place more emphasis on data controllers being accountable for and transparent about their lawful basis for processing. It will change the legal basis that researchers at universities can use to process personal data for research purposes. More specifically, universities will be classified as public authorities for the purposes of this new law and academic researchers will no longer be able to use legitimate interests as their legal basis for processing personal data for their research. The Information Commissioner's Office (ICO), whose role is to ensure compliance with this legislation, further advises that public authorities should avoid relying on consent unless they are confident they can demonstrate it is freely given. Instead, they advise for public authorities including universities to use the 'task in the public interest' basis for everything except their 'internal' functions (e.g. Human Resources processing). At the time of writing this report, the Bill was being discussed in the House of Lords.

The **Digital Economy Act (DEA) 2017** (Part 5, Chapter 5 "Sharing for Research Purposes") provides public authorities with a new power to disclose for research information held in connection with their functions, subject to a number of data security conditions and compliance with the Data Protection Act. The DEA 2017 aims to facilitate the sharing and linking of administrative data held by public authorities to broaden the capacity of research to deliver a number of direct and indirect public benefits. The Act creates a permissive gateway to enable public authorities to make information available to researchers for research that is in the public interest. The Act identifies the UK Statistics Authority (UKSA) as the body for setting the Code of Practice and for accrediting data processors, researchers and research. The Act excludes health and social care data. Devolved administrations and in particular Scotland have additional data related legislations (e.g., the Public Records Scotland, 2011). The Scottish Data Linkage Framework has seven guiding principles that aim to guide decisions on data sharing and linkage for the benefits of research for public and patient good whilst maintaining privacy.

Finally, the **Health and Social Care Act 2012** allows for the dissemination of health and/or adult social care data for research which relates to the provision of health and/or social care, or the promotion of health. Where identifiable data are necessary for research, approval is needed under section251 of the NHS Act 2006, but this must be for 'medical purposes. The Act included restrictions on how the Health and Social Care Information Centre (HSCIC), now NHS Digital, can share data.

In order to provide a firm legal basis and increase public trust in how confidential patient data is being used, in November 2014, the Department of Health appointed Dame Fiona Caldicott as the first **National Data Guardian (NDG)** for Health and Care. In 2016, **the Caldicott Review of Data Security, Consent and Opt-out** was published which proposed measures to strengthen the security of health and care information and help people make informed choices about how their data is used. Since then the UK Government has responded to accept these recommendations, and plans to implement a new national opt out in 2018.

Annex H: Data ownership in the UK

A.W. Boyd

ALSPAC, Population Health Sciences, Bristol Medical School, University of Bristol
CLOSER, Institute of Education, University College London

Data ownership in the UK is complex, with a wide range of potential owners and also where some datasets cover the whole of the UK, some are devolved to the four home nations and some are regional or linked to specific geographies (e.g. health geographies). This wide range of ownership presents substantial challenges to longitudinal studies in negotiating access to these records. It is important to note that there can be substantial differences between the owner of the data and the organisation which acts as the Data Controller of the records (i.e. that studies may need to negotiate access to records with an organisation that is not the ultimate owner of the records).

The primary social administrative data sets of interest to longitudinal studies are owned by government departments. Where: benefits records are owned by the Department for Work and Pensions; employment and earnings records are owned by MH Revenue and Customs; and, criminal records and convictions records are collated within the Police National Computer database but access is controlled by the Ministry of Justice. Ownership of education records is complex, with state maintained English education records to age 16 (and in some cases 18) being collated in the National Pupil Database (NPD, Department for Education) and higher education records being collated in the Higher Education Statistics Agency dataset (which is a charitable company operating under a statutory framework). Education data is devolved to the home nations, with Scotland, Wales and Northern Ireland maintaining similar systems to the English NPD. Privately provided education records are collated by the charitable establishments providing the education, but some records are fed into the NPD system. Ownership of health records are exceedingly complex and reflect operational divisions within the Department for Health. There are substantial differences in ownership over the four home nations of the UK. Primary care data is owned at an individual practice level (which are typically private companies). Secondary care data is collated locally (within operational health geographies) although it is mandated that some information is centralised where access is controlled by NHS Digital. Public Health England also control access to some data as do the Medicines & Healthcare products Regulatory Agency (MHRA) and the NHS National Institute for Health Research (NIHR). For example, the Central Practice Research Datalink (CPRD) is MHRA and NIHR run, but also populated by data originating from General Practice and NHS Digital. Certain health data are collated and controlled within the University sector, for example the NICOR cardiac register is owned by University College London. Ownership of natural environment records is similarly fragmented, with some being owned by local government and fed into government departments (e.g. air pollution monitoring records are managed by Department for Environment, Food and Rural Affairs), some which are managed by Executive Agencies (e.g. the Met Office collate climate and meteorological data), and some of which are owned by research councils (e.g. the National Radon Dataset, managed and produced by the British Geological Survey, is owned by the Natural Environment Research Council). Commercial datasets are owned by a wide range of companies.

Annex I: Current data linkage initiatives in Brazil, Canada, Australia and New Zealand

Ray Chambers and Natasha Codiroli McMaster

Current data linkage initiatives in Brazil, Canada, Australia and New Zealand

As highlighted in the Administrative Data Taskforce (ADT) report in December 2012, the use of linked administrative data in the UK is lagging behind those in Nordic countries and some parts of Europe such as the Netherlands, where advantage has been taken of national datasets based on routinely collected government administrative data. National legislations in the Nordic countries have allowed for a single unique identifier across all administrative datasets and with appropriate data infrastructures in these countries, record data linkage has flourished.

However, progress is currently being made worldwide for more countries to exploit their rich sources of administrative government data. Outside the UK and Europe, the examples that follow are indicative of such efforts.

Brazil

The Brazil 100 million project aims to build a cohort through administrative data of all individuals registered in the CadastroÚnico (CADU) database, which includes around one in three Brazilian residents. The database includes Brazilian households that earn less than half the minimum wage per person. The cohort will include those who have received a grant from the Bolsa Familia (PBF) program (a form of welfare payment) between 2003 and 2015, ultimately resulting in the inclusion of around 114 million individuals.

The project was initially conceived as part of collaboration between Brazil and the UK in 2013. For each individual there are around 5,000 variables. These include hospitalisations, notifiable diseases, births and deaths, and a number of specific disease diagnoses. The dataset will support large research projects into Malaria, Zika and microcephaly, and will be used for genomics studies.

Long-term goals include developing innovative probabilistic linkage methods. In 2016 the Centre for Data Interrogation and Knowledge for Health (CIDACS) was set up to support the linkage activity and resulting research projects. There are also plans to test machine-learning methods for more accurate data linkage.

Canada

Canada has a number of units that are members of the International Population Data Linkage Network (IPDLN), outlined in Table 1. In response to issues with survey response rates, Statistics Canada aim to either replace or complement survey data through use of linked administrative data. There are also long-term plans to replace the Census with high quality linked records.

Statistics Canada has been involved in a number of innovations in data linkage, including the development of software to facilitate accurate linkage, for example G-Link (formerly Generalized Record Linkage System (GRLS)), which is used to perform probabilistic linkage. They have also created a Social Domain Linkage Environment (SDLE), which holds multiple linked administrative datasets including Census, tax filers, births, deaths, landed immigrants, and the Indian registry. The data is held in two components for ease of use and to maintain anonymity, including a record depository which contains identifiers which can be used to link to various records.

There is also on-going work to link survey data with administrative records across Canada. For example, work to link two nationwide longitudinal cohort studies, the Canadian Longitudinal Study on Aging (CLSA) and the Canadian Partnership for Tomorrow Project (CPTP) with Administrative health data (AHD).

Table 1: Data linkage initiatives in Canada

Data linkage unit	Key features
The Child and Youth Data Lab (CYDL)	Managed by the Alberta Centre for Child, Family and Community Research, CYDL links anonymised government data across ministries.
Population Data BC	Population Data BC is a data-linkage and educational resource aimed at facilitating research into human health, wellbeing and development. Three university partners, the University of British Columbia, the University of Victoria and Simon Fraser University, run it. It aims to provide research access to linkable data from sources including health, education, early childhood development, workplace and the environment.
The Institute for Clinical Evaluative Science (ICES)	The ICES is a registered independent charity that maintains a database of health and population data in Ontario.
Manitoba Centre for Health Policy (MCHP)	Conducts health research for Manitoba. Maintains a de-identified database of health, education and human services data.
Statistics Canada	Following the Statistics Act (1918) the Canadian government is required to collect, compile, analyse, abstract and publish statistical information. Most recently, this includes linking administrative data for analysis.

Australia

The Population Health Research Network (PHRN) located in Perth is a nationwide initiative aimed at creating an infrastructure to support appropriate collection and use of linked data. The primary use of this network has been the linking of health and administrative data, however new work streams focusing on education, crime and child protection have emerged from the work. Within the network are a number of data linkage units and the majority of these hold linked-data from specific regions.

One exception is the nationally linked data held by the Australian Institute of Health and Welfare (AIHW), aiming to facilitate analysis that takes account of an individuals' life course and a wider range of policy areas. They have linked both internally held datasets for research purposes including cancer, death and diabetes databases, with externally held datasets including information on benefits receipt, health registries, births, deaths, marriage notices and the electoral role, amongst many others.

An example of a regional linked database that has been particularly successful is the work by the Western Australian Data Linkage Branch, which covers the total population of Western Australia (around 2 million people) and has up to 40 years of records. The branch was established in 1995 as part of a collaboration between Western Australia Department of Health, the University of Western Australia, Telethon Kids Institute and Curtin University. The unit has pioneered new data linkage techniques to help overcome the fact Australian's do not have one identification number, but separate ones for their Medicare, Driver's licence and hospital records, and now has linked over 30 administrative records. By 2016 the data had been used in over 800 projects, resulting in more than 250 journal publications and 35 graduate research degrees. Going forward, the PHRN is hoping to use learning from the Western Australian Branch to establish similar data linkage units in all Australian states and territories.

Table 2: Data linkage initiatives in Australia

Data linkage unit	Key features
Centre for Health Record Linkage (CHeReL)	Data linkage unit for New South Wales and ACT. Managed by the New South Wales Ministry of Health. A record linkage infrastructure for the health and human services sectors provides access to researchers, health planners and policy makers.
Data Linkage Queensland (DLQ)	Located within the Health Statistics Unit at the Queensland Department of Health.
SA-NT DataLink	Record linkage system for health and human services in South Australia and Northern Territory. Members include the Northern Territory government, and numerous SA government departments, the three SA-based universities, the South Australian Health and Medical Research Institute (SAHMRI), the Cancer Council SA and the Health Consumer Alliance of SA.
Tasmanian Data Linkage Unit (TDLU)	Department of Health and Human Services (DHHS) and the Menzies Research Institute Tasmania (MRI).
Centre for Victorian Data Linkage	Victorian state node of the Population Health Research Network. Funded by the National Collaborative Research Infrastructure Strategy and the Victorian Government Department of Health and Human Services. Provides population wide linked data to researchers, builds sills and capacity in using linked data for research and creates opportunities for researchers to undertake research using the data.
Western Australia Data Linkage Branch	Managed and funded by the Western Australia Department of Health. Includes 30 datasets covering the total population of Western Australia and up to 40 years of records.

New Zealand

Statistics New Zealand is currently in the process of developing an Integrated Data Infrastructure (IDI) that will include linked administrative and survey data at the individual level. Initially, administrative datasets were linked separately, for example the Linked Employer – Employee Data (LEED) that gave researchers the ability to trace occupational outcomes based on earlier characteristics and educational achievements. This project has strong similarities to the UK Longitudinal Education Outcomes (LEO) data initiative, which links educational records to tax and employment information.

Currently the IDI includes information on tax, education, health, justice, births and deaths, visas, and benefits receipt. The dataset not only includes permanent residents of New Zealand, but also people who have ever been residents. The data is linked to Inland Revenue records, which act as a 'spine,' and from 2014 includes anyone who has interacted with the Inland Revenue from 1999. Data from the 2013 Census were added in 2015, and plans are underway to link in Ministry of Health records.

High profile examples of the use of the IDI include; research into the long-term outcomes of premature babies at the University of Otago, the creation of a tool to give young people a clear idea about likely career outcomes and pathways depending on their educational choice, and a number of projects aimed at assessing long-term impacts of programmes and policies.

Annex J: An overview of data linkage: Background and methods

Ray Chambers, University of Wollongong

Data linkage is often referred to as record matching or record linkage, and record linkage has long been a strategy for gathering information about the same individual by combining the data for this individual from two or more distinct sources. In this sense, it can be said that record linkage dates back to 2,238 BC, when the Chinese emperor Yao ordered data from a population census and an agricultural census to be visually combined by sorting these separate lists according to the same criteria (e.g. alphabetically), see IBGE (2012). The term was first used in a scientific context in Dunn (1946), where it is defined as a process used to gather facts to compose a "book of life", i.e. a record of events occurring and recordable, experienced by the same individual between birth and death. Since then, dozens of definitions have emerged to describe what we call data linkage. In most cases the term is used in a context where the main objective is to gather information about the same individual or event from different data sources.

According to the Organization for Economic Cooperation and Development “record linkage refers to a merging that brings together information from two or more sources of data with the object of consolidating facts concerning an individual or an event that are not available in any separate record” (OECD, 2006). The UK Administrative Data Research Network adopts the definition proposed by the Task Force created by the Economic and Social Research Council in collaboration with the Medical Research Council and the Wellcome Trust: “Data linkage is the joining of two or more administrative or survey datasets using individual reference numbers/identifiers or statistical methods such as probabilistic matching” (ADRN, 2012, p. 41).

Winglee, Valliant and Scheuren (2005) refer to record linkage as “a process of pairing records from two files and trying to select the pairs that belong to the same entity”. These authors consider record linkage in the context proposed by Fellegi and Sunter (1969), where weights are used to characterise the probability of each pair of records being a true match. In contrast, Winkler (2014) refers to record linkage as “... the science (and art) of matching the same entities (person, business, etc.) using quasi-identifiers such as name, address, date of birth, etc.”, which explicitly allows for potential errors in the linkage result.

Newcombe et *al*. (1959) are usually credited as being the first to formulate record linkage as statistical problem, when they proposed a fully automatic procedure for linking birth and marriage records as part of an investigation of differential fertility in the presence of hereditary disease. Based on their ideas, Fellegi and Sunter (1969) specified a mathematical model for computer-based record linkage that explicitly allows for probabilistic matching. This FS model is now widely used to characterise non-deterministic record linkage procedures.

A drawback of the FS model is its implicit assumption of conditional independence, i.e. linkage errors are independent given the identifying variables used to construct the links. Extensions to the FS model have been proposed, mainly based on the use of alternative algorithms for more accurate estimation of model parameters used to classify record pairs as match or non-match; as well as refinements to the comparison functions used in linkage, making it possible to calculate partial agreement scores when the variables whose values are being compared are subject to variation (Christen, 2007).

Different approaches to non-deterministic record linkage have emerged in the last twenty years. Many of these do not assume conditional independence and are based on techniques that were originally developed for data mining and information retrieval, employing methods that are closely related to those used in machine learning and artificial intelligence (Christen, 2007). In particular, these techniques use supervised learning and so require training data, i.e. a list of correctly linked record pairs (often referred to as gold standard linked data) that can be used to identify patterns of true match and true non-match in the databases that are being linked. In effect, the linkage problem is reposed as a classification problem.

Linkage based on the use of such classifiers has been shown to be superior to linkage based on the FS model (Christen 2007, 2008a and 2008b; Elfeky et al., 2002 and 2003; Liu at al., 2003). However, these classification-based linkage techniques depend on the availability of identifying information that can be used to define a “good” classification model. This can be difficult to achieve in many practical situations. Some alternatives exist (Christen, 2007, 2008a 2008b; Yu et al., 2004) but there is no compelling evidence that they lead to better linkage outcomes.

Most record linkage errors are due to discrepancies in the identifying data for the same individual that are held on the databases being linked, and occur because the linkage method cannot deal with this “fuzziness”. Except for very basic information collected in birth and death registers, most population registers contain non-compulsory information, obtained in different ways and at different times, but ostensibly relating to the same event. The use of this information as identifying data for linkage can then raise concerns about the quality and the coverage of the resulting linked dataset.

In general, record linkage outcomes can be of four types: true positives, when records from the same individual are compared and are correctly identified as such; false positives, when records from different individuals are compared and are said to belong to the same individual; true negatives, when records from different individuals are compared and are said to not belong to the same individual; and false negatives, when records from the same individual are compared and are said to not belong to the same individual. Linkage errors can be due to false positives and false negatives. According to Coeli (2015) “False positive errors are more frequent when few fields are available for comparison, completeness of identifiers is low, the proportion of homonyms is high and linked databases have a high volume of data. False negative errors, on the other hand, happen due to incorrect information, typographical errors and the absence of records of the events in the databases”.

In most non-deterministic linkage methods a trade-off has to be made between choice of the matching threshold used in the comparison function and the resulting number of false positives and false negatives, since these errors move in opposite directions depending on choice of this threshold. In particular, increasing the threshold can lead to fewer false positives but many more false negatives, while decreasing the threshold can lead to fewer false negatives but many more false positives. For example, when linkage is used for building a list for identifying non-vaccinated cases in a population, false positives may be more acceptable than false negatives, and so a lower threshold may be reasonable. On the other hand, if the linkage is being carried out to identify tax evasion cases then false negatives may be more acceptable than false positives and a higher matching threshold needs to be set.

Evaluating linkage quality is of crucial importance to ensure that the limitations of these data are taken into account during analysis. This will depend on available paradata for the implementation of the linkage procedure, and particularly whether these data contain information about the potential distribution of linkage errors and whether this distribution deviates from what would be considered as being “at random”. A number of studies have shown that linkage errors can differ according to demographic characteristics, such as sex and age; place of birth or occurrence; ethnic origin; economic group; year of occurrence etc. (Bohensky, 2016). When linkage errors are not distributed “at random”, any analysis of the linked data should be aware of the potential for bias because the linked dataset may no longer be representative of the underlying population of interest.

Unfortunately, information on the linkage process that would allow calculation of quality measures based on the characteristics of the distribution of linkage errors is often not available. As a consequence indirect assessment may be the only option. In this context, comparison of distributions of variables defined on the linked dataset with corresponding distributions on the original source data (which can be population register data or representative survey data) is a good first step. Another is to compare estimates based on linked data with actual values provided by official or well-accepted sources.

In some cases information on the true linkage error status of compared records may be available or can be obtained in some way. Here a gold standard linked dataset can be created, and the actual quality of linkage can be assessed using statistics based on the number of true positives, false positives, true negatives and false negatives in the gold standard linked data, e.g. via a confusion matrix (Christen, 2012; Hand and Christen, 2017).

Finally, a clerical sample audit or some form of post-linkage review may be possible in order to determine the quality of linkage. The feasibility of this strategy depends on access to the original datasets underpinning the linked dataset and the availability of resources for carrying out the audit/review. As a consequence it may not be feasible for agencies without access to the original data or with access to limited resources for audit/review. However, it is a strategy that should be seriously considered as part of any data linkage exercise, since in many situations estimates of the distribution of linkage errors derived from quite small audit samples may be very informative for assessing the quality of a linkage process.

References

Administrative Data Taskforce (2012). The UK Administrative Data Research Network: Improving Access for Research and Policy.

Bohensky, M. (2015) Bias in data linkage studies, in Methodological Developments in Data Linkage (eds K. Harron, H. Goldstein and C. Dibben), John Wiley & Sons, Ltd, Chichester, UK. doi: 10.1002/9781119072454.ch4

Christen, P. (2007). A Two-Step Classification Approach to Unsupervised Record Linkage. Conferences in Research and Practice in Information Technology - CRPIT, 70, 111-119.

Christen, P. (2008a), Automatic Training Example Selection for Scalable Unsupervised Record Linkage, Pacific Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2008), ed. Takashi Washio, Einoshin Suzuki, Kai Ming Ting, Akihiro Inokuchi, Springer, New York, pp. 511-528.

Christen, P. (2008b), Automatic Record Linkage using Seeded Nearest Neighbour and Support Vector Machine Classification, ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2008), ed. Conference Program Committee, Association for Computing Machinery Inc (ACM), New York USA, pp. 151-160.

Christen, P. (2012). Data Matching - Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection. Springer-Verlag Berlin Heidelberg. DOI: 10.1007/978-3-642-31164-2

Coeli, C. M. (2015). We must pay more attention to record linkage quality. Cad. Saúde Pública, 31(7). http://dx.doi.org/10.1590/0102-311XED010715

Dunn, H. L. (1946). Record Linkage. Am J Public Health Nations Health, 36(12), 1412-6.

Elfeky, M. G., Ghanem, T. M., Verykios, V. S., huwait, A. R. & Elmagarmid, A. K. (2003) Record Linkage: A Machine Learning Approach, A Toolbox, and A Digital Government Web Service. Computer Science technical Reports. Paper 1573. https://docs.lib.purdue.edu/cstech/1573/

Elfeky, M. G., Verykios, V. S. & Elmagarmid, A. (2002). TAILOR: a record linkage toolbox. 17-28. 10.1109/ICDE.2002.994694.

Fellegi, I.P., and Sunter, A.B. (1969). A theory for record linkage. Journal of the American Statistical Association, 64, 1183-1210.

Hand, D. & Christen, P. (2017), A note on using the F-measure for evaluating record linkage algorithms' Statistics and Computing, pp. 1-9.

IBGE (2012). Sínteses Históricas, Históricos dos Censos, Panorama introdutório. Brasil, Ministério do Planejamento, Orçamento e Gestão, *Instituto Brasileiro de Geografia e Estatística*.

Liu, B., Dai, Y., Li, X., Lee, W.S. & Yu, P.S. (2003), Building text classifiers using positive and unlabeled examples, in 'IEEE International Conference on Data Mining' (ICDM'03), Melbourne, Florida, pp. 179–186.

Newcombe, H. B., Kennedy, J. M., Axford, S. J. & James A. P. (1959), Automatic Linkage of Vital Records. Science, 130, 954-959.

OECD (2006, January 4). Glossary of Statistical Terms: Record Linkage. Retrieved from https://stats.oecd.org/glossary/detail.asp?ID=3103

Winglee, M., Valliant, R. & Scheuren, F. (2005). A Case Study in Record Linkage. Survey Methodology, 31(1), 3-11.

Winkler, W. E. (2014), Matching and record linkage. WIREs Comput Stat, 6: 313–325. doi:10.1002/wics.1317

Yu, H., Han, J., Chang, K. and Chen C. (2004). PEBL: Web page classification without negative examples. *IEEE Transactions on Knowledge and Data Engineering*, 16, 70-81.

Annex K: CLOSER: a brief view on consent in longitudinal research

Andy Boyd[1,2], Alison Park[2]

[1] ALSPAC, Population Health Sciences, Bristol Medical School, University of Bristol
[2] CLOSER, Institute of Education, University College London

Seeking participant consent for the use of their information or biological samples is a cornerstone of study ethico-legal frameworks. Neither data protection legislation nor Common Law relating to confidentiality has precisely defined what constitutes 'valid' consent. The imminent EU General Data Protection Regulations (GDPR) will change this:

"*'consent' of the data subject means any freely given, specific, informed and unambiguous indication of the data subject's wishes by which he or she, by a statement or by a clear affirmative action, signifies agreement to the processing of personal data relating to him or her*;" GDPR Article 4 (11).

This reinforces current expectations that consent should be explicit and documented; an existing legal requirement in some areas (e.g. Human Tissue Act 2004). However, the legal necessity for consent is possibly misunderstood, as consent is typically not the appropriate legal basis for studies to meet Data Protection legislation requirements[1].

Within the field of record linkage, seeking and maintaining consent introduces substantial challenges. Prospectively, studies can 'consent by design' at recruitment (e.g. UK Biobank required participant consent for linkage to health records as a precondition to enrolment). Seeking consent within existing studies results in incomplete response; where response rates vary by socio-economic and potentially health status, and consent rates are context specific and vary according to data source, meaning a potential loss of statistical power and introduction of response bias.[2],[3] Changing expectations regarding the 'validity' of consent is proving problematic for longitudinal studies; with consent wording being perceived as 'outdated' or the intervals between 'refreshing' consent being considered overly long. In practice, this is resulting in demands that studies renew consent, which is resource intensive for studies and burdensome and potentially confusing for participants. To meet these challenges, some longitudinal studies make use of consent alternatives (e.g. using 'Section 251' (NHS Act 2006) powers to meet Common Law requirements when accessing identifiable health records without consent) or by adopting 'effectively anonymous' data processing pipelines.

It is tempting to speculate that, if the consent 'gold standard' is not fixed, then it is within the domain of the stakeholders involved to change the way in which best practice is framed. Possibilities emerging from the research derogations to the UK GDPR implementation and initiatives such as the National Patient Guardian's recent report[4] into sharing health records may challenge the status quo position expecting opt-in explicit consent. These changes may introduce opportunities for longitudinal studies, although studies will need to work with participants and the wider public (e.g. through the Understanding Patient Data programme[5] and study based qualitative research[6]) to understand and accommodate their expectations and ensure that public acceptability is maintained while realising these opportunities.

[1] See: https://iconewsblog.org.uk/2017/08/16/consent-is-not-the-silver-bullet-for-gdpr-compliance/

[2] Kho ME, Duffett M, Willison DJ, Cook DJ, Brouwers MC. Written informed consent and selection bias in observational studies using medical records: systematic review. Bmj. 2009 Mar 12;338:b866.

[3] Knies G, Burton J, Sala E. Consenting to health record linkage: evidence from a multi-purpose longitudinal survey of a general population. BMC health services research. 2012 Mar 5;12(1):52.

[4] Review of data security, consent and opt-outs. (2016) National Patient Guardian. Available from: https://www.gov.uk/government/publications/review-of-data-security-consent-and-opt-outs

[5] See: www.understandingpatientdata.org.uk

[6] Audrey S, Brown L, Campbell R, Boyd A, Macleod J. Young people's views about consenting to data linkage: findings from the PEARL qualitative study. BMC med res methodol. 2016; 16(1)34.

Annex L: Analysis of data downloads and publications of the CLS cohorts and Understanding Society

ESRC Office, November 2017

This annex contains information on data downloads and publications for Understanding Society (and the British Household Panel Survey) and the cohorts held at the Centre for Longitudinal Studies (CLS) for the years 2007 through 2016.

1. Publications

Data on publications were obtained from CLS and Understanding Society. Though there is substantial similarity in what information on publications is collected by both studies, the differences in publication type included in the total have been listed below. Both institutions note that the numbers included in this annex are likely an undercount because authors do not always cite the data used and data users do not always report their publications to UKDS or the studies.

Publication figures include the publication types below:

Understanding Society & BHPS	CLS
Peer-reviewed journal articles	Peer-reviewed journal articles
ISER & Understanding Society Working Papers	Peer-reviewed working papers
Books and book chapters	Books and book chapters
Government reports and parliamentary papers	Government and third sector reports
Theses	PhD theses
Research papers	Conference presentations

Table 1a: The number of publications using data from the three national birth cohorts* at the Centre for Longitudinal Studies and from Understanding Society and the British Household Panel Survey (2007-2016)

Publications (2007-2016)												
Year	NCDS		BCS70		MCS		CLS (Total **excl.** Next Steps)*		Understanding Society**		BHPS	
2007	103	6%	31	7%	40	6%	174	6%	n/a	n/a	188	12%
2008	129	8%	36	8%	46	7%	211	8%	5	1%	213	14%
2009	163	10%	45	10%	39	6%	247	9%	3	1%	202	13%
2010	188	12%	54	11%	86	13%	328	12%	6	1%	207	13%
2011	193	12%	42	9%	61	9%	296	11%	27	5%	121	8%
2012	211	13%	57	12%	74	11%	342	12%	53	9%	109	7%
2013	175	11%	45	10%	75	11%	295	11%	75	13%	142	9%
2014	141	9%	54	11%	81	12%	276	10%	119	20%	147	9%
2015	142	9%	53	11%	81	12%	276	10%	131	22%	105	7%
2016	165	10%	55	12%	87	13%	307	11%	175	29%	115	7%
TOTAL	1610		472		670		2752		594		1549	

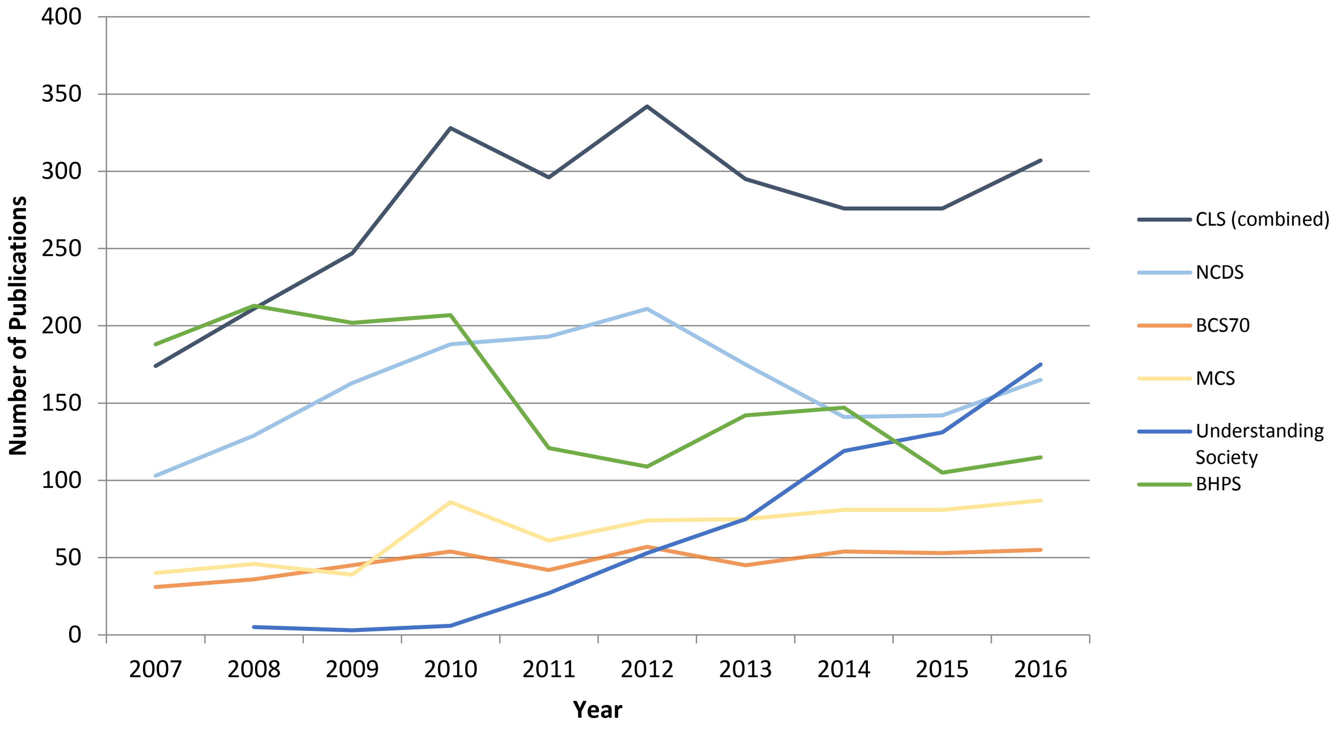
Note: Percentages show the proportion of downloads in the given year compared to the grand total 2007-2016 for that study. Percentages are rounded to the nearest whole number.

Due to differences in how CLS and Understanding Society handle combined studies publications, the figures may not be directly comparable.

*Information on publications from the Next Steps cohort was not available at the time of data collection.

** Understanding Society began in 2009; however, the first *Innovation Panel* began in 2008. Understanding Society data became available from the UK Data Service in 2010. Some publications are included in both BHPS and Understanding Society lists, so the figures cannot be combined.

Figure 1b: The number of publications for CLS, Understanding Society and BHPS, 2007-2016.



2. UKDS Downloads

Data on the number of downloads of data for each study and details of the users downloading the data were obtained from the UK Data Service (UKDS).

The number of downloads serves as an indicator of data access and is a proxy for use; however, a download does not directly link to actual use of the data. The same researcher could download the same dataset multiple times or one download could be used for multiple analyses. Downloads also may or may not result in publications or inclusion in reports.

Download figures for each study contain the following types of download/data access level:

- Centre for Longitudinal Studies (CLS) – all normal downloads, orders and secure access for all four cohorts held at CLS: 1958 National Child Development Study (NCDS); 1970 British Cohort Study (BCS70); Next Steps; Millennium Cohort Study (MCS)
    - A highlight has been placed on MCS, which corresponds to the highlight on MCS in Section 1 of the Longitudinal Studies Strategic Review main document
- Understanding Society– all normal downloads, orders and secure access

- British Household Panel Survey (BHPS) –all normal downloads, low-level geography orders, med-level geography orders, med-level geography non-UK orders and secure access

Requests for access to sensitive data through METADAC have not been included in the UKDS download figures, but are shown in separate tables in Section 3 of this Annex.

Table 2a: The number of downloads for the CLS cohorts, Understanding Society and BHPS data for the period 2007-2016.

Downloads (2007-2016)														
Year	NCDS		BCS70		Next Steps		MCS		CLS (Total **incl.** Next Steps)		Understanding Society		BHPS	
2007	1322	7%	1055	5%	119	3%	361	2%	2857	5%	n/a	0%	2482	8%
2008	1219	7%	1167	6%	229	6%	732	5%	3347	6%	n/a	0%	2463	8%
2009	1322	7%	1296	7%	287	7%	546	4%	3451	6%	n/a	0%	2827	9%
2010	1570	9%	1662	9%	268	7%	900	6%	4400	8%	51	0%	2632	9%
2011	1464	8%	1631	8%	456	12%	900	6%	4451	8%	706	5%	2852	9%
2012	1767	10%	1873	10%	553	14%	1394	10%	5587	10%	1399	9%	2945	10%
2013	1700	9%	1608	8%	429	11%	1359	9%	5096	9%	1799	12%	3020	10%
2014	1806	10%	2154	11%	417	11%	1967	13%	6344	11%	2754	18%	3002	10%
2015	2139	12%	2110	11%	385	10%	2265	15%	6899	12%	2794	18%	3113	10%
2016	2217	12%	2964	15%	383	10%	2259	15%	7823	14%	3318	22%	3065	10%
TOTAL	16526		17520		3526		12683		50255		12821		28401	

Note:

- This covers the period between January 2007 and December 2016, apart from Understanding Society, for which downloads began in December 2010.
- The download count for CLS is the sum of downloads for all four cohorts, including Next Steps.

Figure 2b: The number of UKDS data downloads for CLS, Understanding Society and BHPS, 2007-2016.

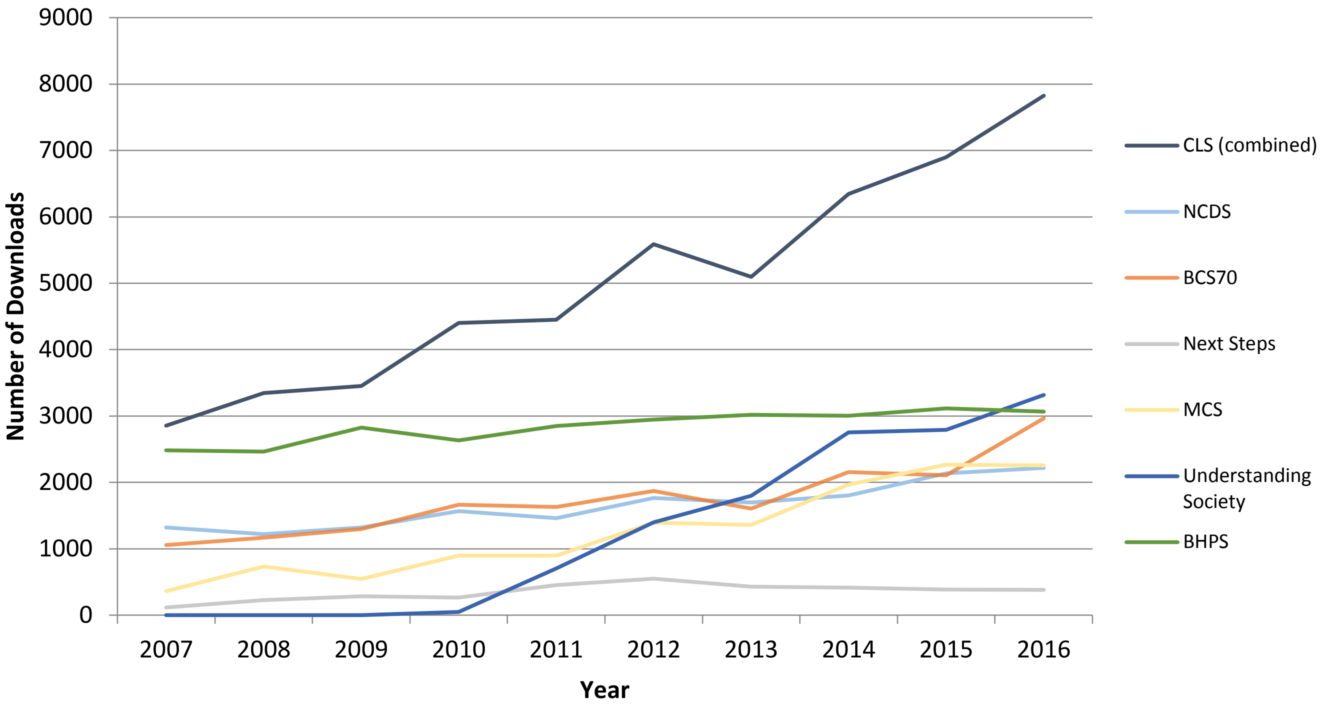


Table 2c: The category of user who downloaded data from UKDS for the CLS cohorts (with highlight on MCS), Understanding Society and BHPS (2007-2016)

User Type	CLS		MCS		Understanding Society		BHPS	
Postgraduate	19152	38%	5405	43%	4532	35%	11695	41%
Staff at institute of HE	19762	39%	5360	42%	4740	37%	6974	25%
Undergraduate	7403	15%	703	6%	1709	13%	6494	23%
Student in further education	1650	3%	507	4%	277	2%	1298	5%
Central Government staff	486	1%	195	2%	614	5%	626	2%
NGO or registered charity staff	683	1%	201	2%	383	3%	440	2%
Other not-for-profit	413	1%	158	1%	229	2%	316	1%
Commercial user	157	0.3%	36	0.3%	145	1%	197	1%
School student	210	0.4%	45	0.4%	13	0%	138	0%
Local Government staff	100	0.2%	29	0.2%	85	1%	69	0%
Personal / genealogical user	103	0.2%	11	0.1%	57	0%	55	0%
School teacher	63	0.1%	11	0.1%	8	0%	44	0%
Staff, other	54	0.1%	20	0.2%	15	0%	16	0%
Research Council staff	9	<0.1%	1	<0.1%	12	0%	16	0%
Academic visitor	0	0%	0	0%	0	0%	4	0%
No data	10	<0.1%	1	<0.1%	2	0%	19	0%
Total	50255		12683		12821		28401	

Note:

- This covers the period between January 2007 and December 2016, apart from Understanding Society, for which downloads began in December 2010.
- CLS data includes Next Steps.
- The UKDS categories "Staff at institute of further education" and "Staff at institute of Higher Education" have been combined into one "Staff at institute of HE" category.

Table 2d: The proportion of data downloads from UKDS by discipline of researcher for CLS (with highlight on MCS), Understanding Society and BHPS (2007-2016)

Study	Discipline	
CLS	Economics and Econometrics	38.1%
	Sociology	13.2%
	Psychology	10.3%
	Other Studies and Professions Allied to Medicine	8.9%
	Statistics and Operational Research	6.4%
	Education	4.8%
	Social Policy and Administration	4.3%
	Community-based Clinical Subjects	2.1%
	Geography	1.3%
	Business and Management Studies	1.3%
	Library or Data/Information Centre	1.2%
	Politics and International Studies	1.1%
	Anthropology	1.0%
MCS	Economics and Econometrics	24.9%
	Psychology	15.6%
	Sociology	13.1%
	Other Studies and Professions Allied to Medicine	12.9%
	Statistics and Operational Research	7.3%
	Education	7.2%
	Social Policy and Administration	7.0%
	Community-based Clinical Subjects	2.1%
	Hospital-based Clinical Subjects	1.3%
	Geography	1.3%
	Social Work	1.1%
Understanding Society	Economics and Econometrics	33.5%
	Sociology	21.1%
	Statistics and Operational Research	8.4%
	Social Policy and Administration	7.2%
	Geography	6.5%
	Other Studies and Professions Allied to Medicine	4.0%

	Psychology	3.8%
	Politics and International Studies	3.0%
	Business and Management Studies	3.0%
	Education	1.5%
BHPS	Economics and Econometrics	57.3%
	Sociology	13.5%
	Social Policy and Administration	4.6%
	Statistics and Operational Research	3.9%
	Business and Management Studies	3.4%
	Geography	3.1%
	Politics and International Studies	2.8%
	Accounting and Finance	2.1%
	Psychology	1.6%
	Other Studies and Professions Allied to Medicine	1.3%

Note:

- This includes disciplines with a 1% or more share of the total downloads for that study.
- This covers the period between January 2007 and December 2016, apart from Understanding Society, for which downloads began in December 2010.
- CLS data includes Next Steps.

Table 2e: The proportion of data downloads from UKDS by country of researcher for CLS (with highlight on MCS), Understanding Society and BHPS (2007-2016)

Study	Country		Total number of individual countries
CLS	United Kingdom	80.2%	55
	United States	7.1%	
	Germany	2.1%	
	Italy	1.9%	
	Australia	1.3%	
MCS	United Kingdom	82.9%	35
	United States	5.5%	
	Ireland	2.2%	
	Italy	1.7%	
	Germany	1.6%	
	Australia	1.4%	
Understanding Society	United Kingdom	88.0%	42
	United States	2.4%	
	Germany	2.2%	
	Italy	1.5%	
BHPS	United Kingdom	78.6%	64
	United States	5.1%	
	Germany	3.7%	
	Italy	1.6%	
	Netherlands	1.6%	
	France	1.1%	

Note:

- This includes countries with a 1% or more share of the total downloads for that study.
- This covers the period between January 2007 and December 2016, apart from Understanding Society, for which downloads began in December 2010.
- CLS data includes Next Steps.

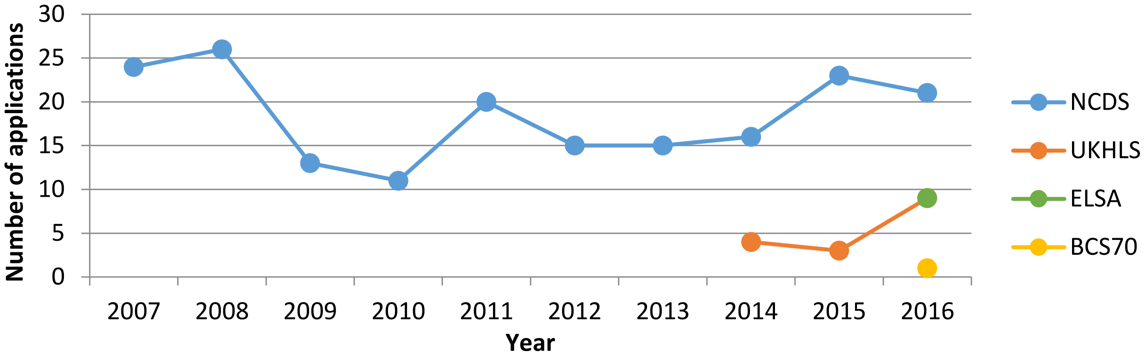
3. Applications for sensitive data

Applications for genotype-survey data and biosamples for Understanding Society and the CLS cohorts NCDS, BCS70 and MCS are accessible by researchers through the METADAC (Managing Ethico-social, Technical and Administrative issues in Data ACcess). METADAC also reviews applications for the English Longitudinal Study of Ageing (ELSA).

Table 3a: The number and outcomes of applications to the METADAC committee (2007-2016) by study.

Decision	Total		NCDS		BCS70		Understanding Society		ELSA	
Approved	184	87.6%	161	88%	1	100%	13	81%	9	100%
Deferred	13	6.2%	4	2%			1	6%		
Declined	5	2.4%	13	7%						
Pending	5	2.4%	1	1%						
Withdrawn	2	1.0%	5	3%						
No data	1	0.5%					2	13%		
Total	210		184		1		16		9	

Figure 3b: The number of applications to the METADAC committee by study for each year 2007-2016.



Note: This includes all applications, not just those approved.

Table 3c: The type of data available through METADAC for each study.

NCDS	DNA
	GWA and exome sequencing datasets (held at the European Genome-phenome Archive)
	Biomedical samples (whole blood, plasma and saliva) and associated biochemical data on certain markers
BCS70	DNA and linked genetic data
Understanding Society	Genetic and epigenetic data (held at the European Genome-phenome Archive; applications for genetic information linked with other data are considered by METADAC)
ELSA	Genetic information linked with other ELSA derived variables

Note: Data obtained from METADAC website and data application forms, January 2018.

Table 3d: The proportion of applications to the METADAC committee (2007-2016) from different countries for total applications and each study individually.

Country	Total		NCDS		BCS70		Understanding Society		ELSA	
UK	152	72.4%	137	74.5%	1	100%	13	81.3%	1	11.1%
USA	28	13.3%	22	12.0%			2	12.5%	4	44.4%
Australia	14	6.7%	14	7.6%						
Netherlands	4	1.9%							4	44.4%
No data	3	1.4%	2	1.1%			1	6.3%		
Germany	3	1.4%	3	1.6%						
Spain	2	1.0%	2	1.1%						
Canada	1	0.5%	1	0.5%						
France	1	0.5%	1	0.5%						
Israel	1	0.5%	1	0.5%						
Switzerland	1	0.5%	1	0.5%						
Total	210		184		1		16		9	

Table 3e: The disciplines of applicants and co-applicants to the METADAC committee.

Discipline	Count	
Epidemiology	143	16%
Chemistry	125	14%
Biosocial studies	103	11%
Genetics	101	11%
Psychology	62	7%
Social Science	45	5%
Sociology	40	4%
Clinical epidemiology	34	4%
Public health	27	3%
Medical Genetics	24	3%
Biostatistics	22	2%
Epigenetics	16	2%
Economics	16	2%
Social work	16	2%
Health Sciences	13	1%
Molecular Genetics	13	1%
Paediatrics	11	1%
Psychiatry	11	1%
Education	11	1%
Oncology	10	1%
Environmental Studies	10	1%
Neuroscience	8	1%
Omics research	8	1%
Nutrition	5	1%
Econometrics	5	1%
Linguistics	5	1%
Molecular Biology	4	0%
Pharmacy	4	0%
Clinical Science	2	0%
Law	2	0%
Management/Business	2	0%
Political science	2	0%
Computing	2	0%
Immunology	1	0%
Communication Studies	1	0%
Employment and work studies	1	0%
Informatics	1	0%

Note: Counts include those who identify with more than one discipline. Includes applications for ELSA data.