Preparing for the future II:
international approaches to challenges facing the longitudinal population studies

# 11:30 – 12:30 Parallel session A

## A1 Reducing attrition and participant engagement: international perspectives

Main hall | Session lead: Hazel Inskip, Southampton Women’s Survey

Key strategies for minimizing attrition and increasing participant engagement in the PSID and its child and young adult supplements
Kate McGonagle, Panel Study of Income Dynamics

The Panel Study of Income Dynamics (PSID) is a longitudinal household panel study that has collected data on economic, social and health behaviour from a nationally representative sample of U.S. families since 1968. The study follows 4,800 original 1968 panel members and their adult children as they have grown up and formed their own economically independent families. As of 2019, approximately 11,000 families participate in the study. Interview data have been collected annually from 1968 to 1997 and biennially from 1999 through 2019. The primary mode of data collection is via computer-assisted telephone interview (CATI) by interviewers employed by the Survey Research Center (SRC) at the University of Michigan. Wave-to-wave response rates in PSID over five decades have historically exceeded 95%; starting in 2015 these rates began to decline slightly and have required more fieldwork effort to maintain. A variety of strategies to increase the likelihood of successful respondent contact and minimize attrition have been undertaken. These include: database searches prior to study launch of interviewer phone numbers that may be blocked and respondent contact information and phone numbers that are not working; advance study materials sent to respondents informing them that interviewers will be calling, the numbers from which they will be calling, and the goals and scientific value of the study; ongoing reminders (telephone, email, text) to complete the study sent to respondents during data collection with consistent study branding and targeted appeals; a respondent website and between-wave respondent newsletters providing study information; a between-wave mailing requesting updated contact information; offering respondents the use of an online appointment scheduler; mixed-mode (web/CATI/face-to-face) approaches to collect data; and a variety of generous respondent incentive strategies, including respondent-focused time-delimited monetary incentives offering additional payments for completing the interview within specific time frames (including an end-game incentive), informant incentives to locate sample members, and interviewer incentives such as motivational messages, small gifts, and financial incentives. These strategies are designed to minimize attrition and increase the efficiency of fieldwork by making it more likely that respondent contact is made and the interview is completed.

This presentation will describe these approaches and provide results from experimental assessments of an online appointment scheduler and time-delimited incentives on data collection outcomes.

Innovative Protocols and Extensive Local Knowledge: How to Overcome Attrition and Participant Engagement Challenges in Longitudinal Studies
Johanna Choumert-Nkolo and Luca Privinzano, Tazanian Kagera Health and Development Survey

EDI Global is a research house conducting large-scale surveys and impact evaluations in East Africa, with a strong experience in conducting longitudinal studies and tracking surveys.

The Kagera Health Development Survey (KHDS) is one of the longest panel data sets from sub-Saharan Africa. The study focuses on the long-run wealth dynamics of households and individuals within the Kagera region, in north-west Tanzania and there were four rounds of data collection from 1991 to 2010.

This presentation discusses the attrition and participant engagement challenges faced during the many rounds of this longitudinal study. Although these challenges existed, EDI Global were able to track 88% of the original respondents from baseline (1991) to endline (2010). Learning from this success, the presentation also focuses on the strategies used by EDI Global to achieve such a high participation rate. These include innovative field protocols, extensive local knowledge and flexible field teams - all of which can be replicated in other longitudinal studies.

## A2 Data linkage

Tavistock room | Session lead: Sally Bridges, Born in Bradford

An overview of the benefits and challenges to data linkage in Australia
Galina Daraganova, Australian Institute of Family Studies

Policy makers, professionals and researchers are demonstrating increasing interest in linking administrative data from multiple sources to enhance longitudinal surveys. While data linkage provides great opportunities, it is not a trivial exercise and comes with a number of challenges.

This talk will provide an overview of the benefits and issues associated with data linkage for longitudinal surveys, emphasising those specific to Australian context. The issues will cover state/territories vs national linkages, consented vs. non-consented linkages, quality of linked datasets, and biases introduced by linked datasets.

Epicosm: a framework for linking and social media data in large-scale birth cohorts
Oliver Davis, University of Bristol

Online social networks such as Twitter, Facebook and Instagram present a great opportunity for passive collection of frequent ecologically valid measurements from large groups of people such as the CLOSER cohorts. Many different types of information can be derived from social media feeds, and one of the most promising is high-resolution measurement of mental health.

CLOSER work package 21 has been working with CLOSER cohort leaders and two generations of participants in the Avon Longitudinal Study of Parents and Children (ALSPAC) to develop Epicosm: a secure and easily deployed software framework for linking data from social media in longitudinal birth cohorts. Our hope is that the software will not only make it easy for cohort studies to link new high-frequency indicators of mental health to existing datasets, but that it will also provide a unique platform for the validation of information derived from social media against gold-standard traditional measures.

## A3 Data harmonisation

Foster and Bloomsbury room | Session lead: Dara O’Neill, CLOSER

**Maximising comparability: Harmonisation of mental health measures in the British birth cohorts**Eoin McElroy, University of Leicester

Common forms of psychological distress (e.g. depression, anxiety) contribute substantially to the global burden of disease. The British birth cohorts contain rich data on these difficulties, and therefore represent a key resource in our attempts to understand the developmental course of these phenomena, along with changes in trends across generations. However, the instruments used to assess mental health vary substantially both within and across these studies. As such, this project aimed to harmonise the various measures of mental health that are available in five British cohorts: NSHD, NCDS, BCS70, ALSPAC and MCS.

This paper will discuss the two-step process used to construct these harmonised measures. In the first step, two researchers independently identified candidate items from different measures that tapped the same underlying construct (e.g. low mood, fatigue, worry). Tests of inter-rater reliability indicated a high level of agreement between the two researchers. In step-two, latent variable modelling approaches were used to harmonise these candidate items. Challenges, limitations and learnings from this project will be discussed.

Dementias Platform UK (DPUK): Facilitating cross-cohort analysis in a digital age
Sarah Bauermeister, Dementias Platform UK

Dementias Platform UK (DPUK) is a £53M public-private partnership established by the MRC to provide access to large-scale cohort data and accelerate the research and discovery of new treatments for dementia. DPUK facilitates multi-modal data access to 42 cohorts across 3.4M individuals within a remote access data repository, the DPUK Data Portal. Globally, there is the need for a solution to manage, process and curate data which is secure, robust, persistent and auditable. DPUK meets this need by managing datasets that are increasingly large, complex and sensitive which are decreasingly feasible to download, transfer and store. DPUK is also investing in standardisation of cohort data for cross-cohort analysis with C-Surv, the DPUK data curation programme.

Here I present the latest data discovery and curation developments for the DPUK Data Portal and how digital innovation is enhancing cross-cohort and multi-modal analysis.

# 13:45 – 14:45: Parallel session B

## B1 New forms of data collection

Main hall | Session lead: Lisa Calderwood, Next Steps

New forms of data collection: Using apps to measure household spending
Jon Burton, Understanding Society

With the ownership and use of mobile devices increasing, researchers are interested in how we can use these to improve data collection. Using the Understanding Society Innovation Panel, we have experimented with ways to collect information about household spending using mobile devices, through a receipt-scanning app and a spending diary app.

In this presentation I will briefly describe what was done, but more importantly what we learned from our experiences in implementing these new forms of data collection.

Taking phenotypes to the population: The ultimate challenge
Melissa Wake, Generation Victoria

Over the last 30 years, many child and adult problems (like mental health, autism, obesity, and allergy) have worsened, with long-term consequences for ageing societies. Research is demonstrably not keeping pace with need. A major block is the lack of very large cohorts well phenotyped for physical, mental and social characteristics, which could provide the evidence needed for better prediction, prevention, treatment and more targeted and equitable services. This gap reflects the burdensome, slow and costly nature of traditional research participant assessments.

In this presentation, we share our steps towards high-throughput cross-generational digital collection of multiple precision phenotypes at scale, firstly with the Longitudinal Study of Australian Children’s new biophysical module (the Child Health CheckPoint, doi 10.1136/bmjopen-2019-030833) and next with Generation Victoria (GenV), targeting 170,000 babies born in 2021-22 and their parents. By solving the technical problem of how to remotely measure core outcomes relevant to families, services, communities and policy, the next generation of large longitudinal studies could support discovery, trials, registries and health services research. This could achieve lasting change in the landscape of large scale research.

## B2 Data linkage

Tavistock room | Session lead: Andy Boyd, Avon Longitudinal Study of Parents and Children (ALSPAC)

Linking data on proximity to fast foods to the Millennium Cohort Study
Emla Fitzsimons, Millennium Cohort Study

This talk will discuss opportunities for and challenges in linking detailed geo data to longitudinal studies, with a specific focus on distance to fast food outlets for MCS families. It will use this rich longitudinal linked data to study the links between proximity to fast food outlets and obesity, across childhood and adolescence. This offers an important advance on many previous studies linking fast food availability and obesity, which are mainly cross-sectional in nature and use less detailed geo data. The majority of previous study are also focused on adults, yet this is a particularly important question for today’s generation of young people given the high rates of excess weight amongst this population.

Challenges and solutions - linking additional data to the UK Census Longitudinal Studies
Oliver Duke-Williams, CeLSIUS

The three census longitudinal studies have had mixed experiences in linking administrative and other external data sets to the core records. This presentation will look at difficulties that have been faced in trying to link data, and try to draw out lessons from successes.

## B3 Data discoverability in biomedical science

Foster and Bloomsbury room | Session lead: Nic Timpson, ALSPAC

Longitudinal data discoverability with MIDUS and DDI
Barry Radler, Midlife in the United States

In the past two decades, advances in computing technology have made available ever-increasing amounts of information while at the same time providing ever-more practical approaches to efficiently managing those data. The inclusion of comprehensive metadata in the research process greatly clarifies the methods used to capture and produce datasets while providing users of those datasets information needed to better discover, interpret, analyse, and share them. Richly structured metadata are even more important with longitudinal studies that contain thousands of variables of many different types. MIDUS (Midlife in the United States) is a national longitudinal study of approximately 12,000 Americans that studies aging as an integrated bio-psycho-social process.

MIDUS has a broad and unique blend of social, health, and biomarker data collected over 20 years through a variety of modes. For the last decade, MIDUS has increasingly relied on a metadata standard called the Data Documentation Initiative (DDI) to manage and document these complex research data. One result of this approach is the creation of a DDI-based, harmonized data extraction system. Such a system allows researchers to search across datasets for variables of interest, identify and harmonize related longitudinal versions of variables, and easily create customized data extracts and codebooks that are directly related to their research questions. This allows researchers to spend more time analysing data instead of managing, merging, or searching for it.

ATLAS: making the UK's cohorts discoverable
Phil Quinlan, HDR UK Atlas Project

The ability for researchers to find suitable resources for their research remains one of the greatest challenges, regardless of whether this is data or samples. There is an inherent need for these assets to be found in order to unlock potential new research. However, discovery is stifled by two common approaches. The first approach is to set a minimum data standard and all resources present data to that minimum level. The challenge that emerges however is that the richness of some of the UK's best data and sample resources is lost and hidden behind the reduction of detail. The second approach is to undertake large scale and in-depth harmonisation but that can take time before results can be revealed.

The approach of ATLAS has been: (1) Bringing in a technology provider to provide the discovery layer (2) Provide resource to support cohorts in connecting (3) Only undertake lightweight harmonisation. The result is a live service that can now be accessed and we are in a pilot phase of release. ATLAS should not be seen as a competing endeavour to many of the community harmonisation efforts and is one of complementarity as we seek to bring wide discovery of cohorts.

# 15:00 – 16:00: Parallel session C

## C1 Data harmonisation

Main hall | Session lead: Rebecca Hardy, CLOSER

Data harmonisation: The LifeCycle Experience
Angela Pinot de Moira, LifeCycle

The LifeCycle collaboration aims to facilitate the utilisation of data from mainly European and also non-European cohort studies for research. It has a particular focus on preconception, foetal and early childhood stressors and their influence on cardio-metabolic, respiratory and mental health trajectories. To achieve its aims, it has established the EU Child Cohort Network, a sustainable data resource and infrastructure which implements the FAIR principles in order to make each participating cohort’s data findable, accessible, interoperable and reusable.

This exemplar presentation will outline the main challenges faced in harmonising data for the EU Child Cohort Network, and how we have attempted to address these challenges.

These include:

* harmonising longitudinal measures without losing information
* balancing precision and inclusivity
* addressing quality control

Growing Up Healthy across the Globe - exploring harmonisation as a tool to understand what shapes wellbeing across population contexts
Susan Morton, Growing Up in New Zealand

Conducting comparative harmonised analyses of longitudinal studies across socio-political contexts offers both opportunities and challenges. In 2016, the Growing Up Healthy in Families Across the Globe (GUH) project was established with the aim of exploring the feasibility of conducting harmonised analysis across five international longitudinal studies which had been established independently of each other. Seed funding was provided by the New Zealand government to explore the potential for harmonisation of analyses across international cohorts and contexts to enhance the policy relevance of findings from specific New Zealand studies of population wellbeing. Five longitudinal studies - three from New Zealand, and one each from Ireland and Scotland were used to determine the feasibility of this harmonisation approach. The studies were chosen because of the similarity of their conceptual frameworks around wellbeing, but also importantly because of the strength of existing collaborative relationships between the Principal Investigators. The general approach taken, the initial findings and the challenges and opportunities of this exploratory harmonisation will be briefly reviewed.

## C2 Reducing attrition and participant engagement

Tavistock room | Session lead: Andrew Wong, MRC 1946 National Survey of Health and Development

The experience of the Uruguayan Survey of Nutrition, Child Development, and Health (ENDIS)
Christian Beron and Nathalia Martinez, Uruguayan Nutrition, Child Development and Health Survey

ENDIS is the first early childhood study in Uruguay that employs a panel methodology in a national household survey. The ENDIS, implemented by the Government of Uruguay, seeks to generate knowledge to advance early childhood policies in Uruguay based on evidence about children’s nutritional status, development and health, and the changes that occur as they grow.

The presentation will be about the Uruguayan work in reducing attrition and participant engagement of the ENDIS, the challenges and complexities faced as well as the lessons learned.

In it for the long haul: addressing attrition and boosting participant engagement at TwinsUK
Paz Garcia, Twins UK

TwinsUK participants range from 18 to 100 years old, with a mean age of 59. TwinsUK is designed to study ageing and frailty, amongst other conditions, so how do we address the challenge of keeping these individuals engaged as they age and become frailer? In addition, how do we ensure our research programme and our communications are relevant for all of our participants, including the younger adults? Lastly, how do we ensure continued participation over many years?

This talk will discuss TwinsUK strategies, participant incentives and use of technology to reduce attrition and keep participants engaged throughout their adult life.

## C3 Data discoverability: issues in interoperability

Foster and Bloomsbury room | Session lead: Nathan Cunningham, University of Sheffield

Building a consensus on both technical and conceptual interoperability
Jon Johnson, CLOSER

A driving force of CLOSER Discovery has been standards. It is architected around the DDI-Lifecycle model, which focuses on the relationships between different parts of the data design, data collection, and data discovery processes. As important as this is, for technical interoperability with other infrastructures which utilise either DDI-Lifecycle or indeed DDI-Codebook; conceptual (semantic) interoperability is also critical if discovery and comparability are to be leveraged at scale.

The presentation will describe how CLOSER Discovery uses HASSET (a subset of ELSST which is integrated into the CESSDA Data Catalogue) and MeSH (which underlies PubMed) for high level conceptual description; and a range of controlled vocabularies which are widely adopted internationally for survey description and methodology. A discussion point will be that within the boundaries of social and biomedical sciences, whilst it provides a powerful basis for cross domain discovery, is it possible or feasible as a community to cohere around such common standards.

Thus far, the focus of CLOSER Discovery has been retrospective documentation. The next few years could be about how such rich metadata can be reused to improve future data collection, data management and in harmonisation. A discussion point will be - what would need to be put in place, if such a vision is to be achieved, or indeed whether such a vision is possible?

Challenges and successes of CESSDA ERIC
Mari Kleemola, Finnish Social Science Data Archive and CESSDA

One of the goals of CESSDA ERIC is to be leading in data discovery. This presentation will give an insight to how CESSDA has been building a coherent infrastructure in a multilingual environment and with Service Providers with differing capacities. I will also shortly describe the approach to (meta)data interoperability taken by the Social Sciences and Humanities Open Cloud (SSHOC) project that contributes to the Open Science agenda and realising the EOSC.