Conference report

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

February 2020

About CLOSER

The UK is home to the world's largest and longest-running longitudinal studies. CLOSER aims to maximise their use, value and impact both at home and abroad. Bringing together eight leading studies, the British Library and the UK Data Service, CLOSER works to stimulate interdisciplinary research, develop shared resources, provide training, and share expertise. In this way, CLOSER is helping to build the body of knowledge on how life in the UK is changing – both across generations and in comparison to the rest of the world. CLOSER was funded by the Economic and Social Research Council (ESRC) and the Medical Research Council (MRC) from 2012-17, and by the ESRC from 2017 to present. Visit www.closer.ac.uk.

The views expressed in this work are those of the conference delegates and do not necessarily reflect the views of CLOSER, UCL, ESRC, MRC or the Wellcome Trust.

This document is available in alternative formats.

Please contact CLOSER:

tel: +44 (0)20 7612 6938

email: closer@ucl.ac.uk

Contents

Executive summary

In January 2020, representatives of 43 longitudinal population studies attended CLOSER's second collaborative conference, *Preparing for the future II: international approaches to challenges facing the longitudinal population studies* in London, UK. The purpose of this conference was to unearth best practice, and identify ways to tackle shared challenges. The programme was informed by the results of a consultation exercise with UK and international longitudinal population studies, which identified data harmonisation, data linkage, new forms of data collection, data discoverability, and reducing attrition and participant engagement as the most important issues facing the longitudinal community.

The conference report documents the day's discussions, including key challenges and opportunities, as well as knowledge and skills gaps, put forward by delegates. This Executive Summary provides an overview of the key points raised.

1.   Data harmonisation

1.1   Successful data harmonisation on a large scale is a time-consuming endeavour that is only feasible through collaboration and coordination, alongside greater awareness of the importance of data documentation.

1.2   Greater consideration of the use of prospective harmonisation attempts should allow for a less complex and resource-intensive process moving forwards.

1.3   Key challenges facing data harmonisation efforts include a lack of standardisation, both in terms of study design, and the use of instruments, validated scales or measures.

1.4   Technological advances are a cause for optimism with data stored and shared in secure virtual environments.

1.5   Data harmonisation efforts at an international level exponentially increase the challenges involved and require greater consideration of linguistic, cultural and other issues around comparability of data.

2. Data linkage

2.1 There is an urgent need to develop a shared narrative and language around the benefits of data linkage, in order to convince data holders, participants and the public of the importance of carrying out this work.

2.2 Transparency and accessibility of communications materials about the data linkage process can help assuage fears and dispel misunderstandings of what data linkage involves and its risks.

2.3 New opportunities for linkage are constantly emerging – however similar challenges persist around data quality, research utility, consent, and participant and public acceptability.

2.4 Linkage potentially offers opportunities to improve information held on population subgroups, such as minorities and vulnerable people, which can be difficult to retain in traditional research studies. However, more information is needed on the coverage and suitability of data on such groups held within linkable datasets.

2.5 Upskilling users to take advantage of the opportunities afforded by linked data would ensure the investment of time and money in creating these resources have strong returns.

3. New forms of data collection

3.1 Delegates highlighted the need for an effective mechanism to share information on successes, failures, risks and benefits. This should include engaging with not just the longitudinal community but new disciplines, commercial companies, market research experts, funders and study participants.

3.2 With many new forms of data collection, response rates can be low, particularly among certain groups. More feasibility studies are needed to understand what works well, including with specific demographics, and to realise the opportunities new methods can offer.

3.3 The development of new technology demands skills that are not traditionally acquired in longitudinal study teams. Studies should work together and share resources to help reduce the heavy financial and resource burden.

3.4 The use of new forms of data collection should be driven by science, and by utilising the best tool to answer a specific question. It should not be driven by technology and innovation for innovations sake.

4. Data discoverability

4.1 All initiatives aimed at improving discoverability and interoperability must have clear objectives that are based on a strong understanding of the target users' requirements and research needs.

4.2 There is demand for discovery platforms that provide highly detailed metadata, as well as for platforms that provide a higher level of information, and allow users to quickly find out if a certain study or data exist.

4.3 Interoperability in discovery platforms can facilitate interdisciplinary research and reproducibility by facilitating research that combines data across domains, and ensure infrastructure adheres to FAIR data principles.

4.4 Sustainable and interoperable infrastructure must be prioritised in funders' investment strategies, and the contribution of such work to novel research, open science and reproducibility recognised.

4.5 There is a need for a discoverability leadership body with technical skills and knowledge, the ability to advise on standards, and a strong management team capable of coordinating across multiple stakeholder groups.

5. Reducing attrition, and participant engagement

5.1 While monetary incentives are not universally appropriate or affordable, most studies make use of some form of 'reward' for participation.

5.2 Participants' family members remain critical influences on participation throughout the life course, and studies should treat them as allies in participant engagement strategies.

5.3 Cultural and country context deeply affects tracing and participant engagement efforts, both in terms of the barriers presented, and the appropriateness and practical viability of different solutions.

5.4 Survey design can be a key engagement tool – making participation flexible and fun can raise response rates.

5.5 When engaging participants, most studies are keen to move towards a model of tailored communication for different sub-groups, but not all can afford it.

5.6 To improve participant engagement, the longitudinal community needs adequate financial resources and a strong literature base on what works.

About the conference

Longitudinal population studies currently face a number of challenging issues. How should studies integrate novel data collection methods and emerging technologies? How can the utility of historic data be enhanced? And critically, how can studies keep participants involved for the long term?

At the same time, studies across the world are finding ways to harness the opportunities these challenges can bring. Following the success of CLOSER's 2018 *Preparing for the future* conference, we extended the discussions beyond the UK, to studies around the world. The purpose of this conference was to unearth best practice, and identify ways to tackle shared challenges. It was a collaborative, interactive event with a strong focus on problem solving.

The programme was informed by the results of a consultation exercise with UK and international longitudinal studies, which asked study teams to identify and rank the key challenges facing them now and in the future. The following topics were selected from this exercise to form the conference programme:

- Data harmonisation
- Data linkage
- New forms of data collection
- Data discoverability
- Reducing attrition and participant engagement

Each session at the conference involved short presentations of exemplar work in a given area, followed by facilitated group discussions and the chance for delegates to share their own experiences.

In total, 102 delegates attended, including representatives of 43 longitudinal population studies from across the globe:

- Aberdeen Children of the 1950s cohort
- Avon Longitudinal Study of Parents and Children (ALSPAC; Children of the 90s)

- 1970 British Cohort Study (BCS70)
- Born in Bradford (BiB)
- CeLSIUS
- Children Growing Up in Liverpool (C-GULL)
- Dementias Platform UK
- English Longitudinal Study of Ageing (ELSA)
- EU LifeCycle
- Evaluation through Follow-up (EFT)
- The Generations Study
- Generation Scotland
- Generation Victoria (GenV)
- Growing Up in Australia: The Longitudinal Study of Australian Children (LSAC)
- Growing Up in Ireland
- Growing Up in New Zealand (GUiNZ)
- Growing Up in Scotland (GUS)
- Hertfordshire Cohort Study (HCS)
- Healthy Ageing in Scotland (HAGIS)
- Kagera Health and Development Survey
- Longitudinal Aging Study Amsterdam (LASA)
- Macmillan HORIZONS Programme
- Methods of Assessing Perinatal Anxiety (MAP)
- Midlife in the US
- Millennium Cohort Study (MCS)
- Million Women Study
- MRC 1946 National Survey of Health and Development (NSHD)
- 1958 National Child Development Study (NCDS)
- Next Steps
- Netherlands Study of Depression in Older Adults (NESDO)
- Northern Ireland Longitudinal Study (NILS)
- ONS Longitudinal Study
- Panel Study of Income Dynamics (PSID)

- REACH Study
- Soweto First 1000 Days Cohort
- Southampton Women's Survey (SWS)
- Twins Early Development Study (TEDS)
- The ELIPSS Survey
- TwinsUK
- Understanding Society: The UK Household Longitudinal Study (UKHLS)
- Uruguayan Study of Nutrition, Child Development and Health Survey (ENDIS)
- Wirral Child Health and Development Study
- Young Lives

Data harmonisation

Session leads: Dr Dara O'Neill and Prof Rebecca Hardy, CLOSER
Report author: Dr Neil Kaye, CLOSER

Key learning

- Successful data harmonisation on a large scale is a time-consuming endeavour that is only feasible through collaboration and coordination, alongside greater awareness of the importance of data documentation.

- Documentation is pivotal to both the validity of any harmonisation exercise and to the utility of any outputs generated.

- Greater consideration of prospective harmonisation should allow for a less complex and resource-intensive process moving forwards.

- A key challenge facing data harmonisation efforts is a lack of standardisation, both in terms of study design, and the use of instruments, validated scales or measures.

- Data harmonisation efforts at an international level can greatly increase the challenges and require greater consideration of linguistic and cultural differences.

Introduction

Data harmonisation can offer many new research opportunities by facilitating cross-study research and by enabling increased sample sizes through data pooling. Retrospective harmonisation, the process of making data already collected by different sources more comparable, is however both time-consuming and challenging. Diverse methods with varying complexity can be employed to harmonise data from different sources or time periods, from simple recoding to latent variable analyses, yet all involve common challenges, including the resource-intensiveness of such work and the potential need to move towards a point of commonality (a lowest common denominator) that can lead to information loss. Such obstacles can become ever more salient when bringing together data from

different regions or countries due to increased potential for variability in cultural contexts, languages, public policies, geographies and data regulation.

Collaboration, knowledge sharing and coordination of effort are each pivotal to overcoming these challenges. Underpinning all of these is the importance of documentation. Harmonisation as a process can often result in information loss, in terms of the precision and complexity of the source data but also with regards to necessary and important metadata. Effective documentation practice is therefore paramount to any harmonisation exercise, in terms of generating useful outputs and ensuring the learning is usefully retained.

Similarly, increased efforts to achieve prospective harmonisation and to establish banks of calibrated measurements could help avoid some of the obstacles that arise in harmonising data already collected by heterogenous means.

The session chairs put the following questions to delegates for discussion:

- What are the specific challenges of harmonising existing measures/data within your study or in your area of research?
- Can we make data collection protocols and instruments more aligned in the future without losing continuity of measurement within studies or responsiveness to new research needs?
- What additional challenges arise when harmonising data from cross-national studies and how can we better facilitate such international harmonisation efforts?

Summary of themes arising from discussions

Comparability within and between studies

The need for improved comparability across studies and across time was emphasised in the discussion as key to furthering our understanding of important cross-contextual differences. Harmonisation however must be driven by specific research requirements, as there is no catch-all solution and it is a process full of obstacles and challenges.

A central challenge related to a current lack of standardisation, both in terms of study design, and the use of instruments, validated scales or measures. The point was raised about the lack of agreement over standard measures, the importance of differing study needs, and the pertinence of historical decisions made within individual studies. Moreover, it was also debated what impact ever-changing developments in technology, administrative protocols and policy priorities would have.

One solution within studies is to retain the validated instruments and measures as far as possible, but ensuring calibration is undertaken between measures when changes in measurement protocols are required. The use of legacy codes and classifications can also facilitate backward-comparability, creating a 'cross-walk' between later and earlier versions.

Between studies, this clearly necessitates greater levels of resourcing and additional collaboration and oversight. It is important, therefore, that studies also consider the potential for using prospective harmonisation approaches. Nonetheless, these carry their own challenges – to ensure a balance between cross-study (and cross-national) comparability and the context-specific nature of the cultural and policy environment. The potential benefits of feeding harmonisation efforts and learning into the creation of an item/measurement bank to guide future data collection planning were also discussed.

Importance of documentation and discoverability

There was considerable discussion centred around the role of documentation and discoverability, and how these are essential precursors to effective data harmonisation. It was recognised that there is a need for greater engagement between data managers and researchers, to ensure that data being harmonised is adequately understood and that the new data resources generated through these efforts have compatibility with the wider data infrastructure from which they have been sourced to ensure wide and long-term utility. It was argued that documentation can be burdensome yet the loss of pertinent metadata can obscure

the provenance and subvert the ability to validate and extend harmonised datasets, so documentation is pivotal to appropriate and worthwhile harmonisation efforts.

International collaboration and advocacy

A major theme in the discussions was the need for greater international research collaboration and the importance of data harmonisation work in maximising the research potential of such efforts. There was particular enthusiasm for cross-study international work centred on generating policy insights. Whilst most of the data are collected through studies undertaken on a national (or even regional) scale, the exigencies of cross-national research in many areas means that international harmonisation would be hugely valuable.

However, the challenges of data harmonisation can be greatly increased in an international comparative context, such as the various linguistic and cultural differences that can result in measurement differences. Conversely, discussions also acknowledged that such contextual differences may be pertinent and therefore there may be a need to maintain culture-specific information. Ultimately, it was agreed that the relevance and suitability of harmonisation needed to be informed by the specific research purposes underpinning its consideration.

The need for support and coordination at an international level is pressing and the discussants highlighted the importance of identifying a champion of such comparative work that recognises the role and importance of harmonisation in maximising research opportunities and impact. This support was required in addition to the need for greater resources.

Resource and training needs

The availability of resources was a key challenge highlighted by all participants in the discussions. Longitudinal studies generate vast amounts of data, which means that retrospective data harmonisation can be hugely labour intensive. Harmonisation requires a substantial amount of technical and procedural knowledge and needs to be undertaken by researchers familiar with the underlying data and the concepts. There remains a need for additional guidance and training

on approaches to harmonisation and the validation of measures that are generated by such processing.

To this end, it was agreed that further funding would be beneficial for retrospective data harmonisation projects and external calibration studies. It was also agreed that a greater emphasis ought to be placed on collaboration towards more prospective harmonisation efforts, including standardisation of documentation of study metadata as well as documentation of retrospective harmonisation efforts.

Data linkage

Session leads: Sally Bridges, Born in Bradford; Andy Boyd, Children of the 90s (ALSPAC)
Report author: Rob Davies, CLOSER

Key learning

- There is an urgent need to develop a shared narrative and language around the benefits of data linkage, in order to convince data holders, participants and the public of the importance of carrying out this work.
- Transparency and accessibility of communications materials about the data linkage process can help assuage fears and dispel misunderstandings of what data linkage involves and its risks.
- New opportunities for linkage are constantly emerging – however similar challenges persist around data quality, research utility, consent, and participant and public acceptability.
- Linkage potentially offers opportunities to improve information held on population subgroups, such as minorities and vulnerable people, which can be difficult to retain in traditional research studies. However, more information is needed on the coverage and suitability of data on such groups held within linkable datasets.
- Upskilling users to take advantage of the opportunities afforded by linked data would ensure the investment of time and money in creating these resources have strong returns.

Introduction

Linked data have become an integral, rather than adjunct, part of longitudinal population studies. The addition of linked records to survey data can offer incredible opportunities to generate research findings with strong practical and policy implications – and as sources of data grow, so too do the opportunities for meaningful linkages. These opportunities – if well articulated – could be used to

shape understanding of data linkage and its value, ultimately helping to overcome some of the barriers studies currently face in linking data.

Longitudinal population studies continue to struggle with bureaucratic hurdles in carrying out linkages, and can also face problems with management and quality of such data. Across countries, studies' linkage efforts can be hampered by shared challenges, such as a lack of adequate analytical skills to cope with linked datasets, and issues with data quality in administrative and other records. Consent in particular remains a major barrier to linkage, both in terms of gaining informed consent from participants as well as gaining consents that meet data holders' (sometimes changing) requirements. Finally, many studies struggle with a lack of unified access to administrative datasets across different territories within the same country, which can often be costly to obtain.

The session chairs asked delegates to consider:

1. How can longitudinal population studies better articulate the value and process of data linkage to different audiences, including data holders, participants and the public?
2. What are the emerging opportunities for data linkage, including linkage to spatial and social media data?
3. How do we include subgroups – and in particular vulnerable people – in linkage strategies?

Summary of themes arising from discussions

Communicating the value and process of linkage to key stakeholders

Delegates suggested that there is an urgent need to develop a shared narrative and language around the benefits of data linkage, in order to convince data holders and participants of the importance of carrying out this work. This 'story' of data linkage would need to cover the unique advantages of linking to longitudinal survey data over data collected through other designs, and include examples of measurable impact achieved and possible using such linked data. Case studies of impactful research based on linked data would bolster the case.

However, delegates noted that a shared narrative would have to account for differences in attitudes to data linkage between countries, and between territories within the same country. It would also need to be tailored to the different perspectives of participants, data owners, and government. While there was consensus among studies that speaking with one voice on this issue was important, it was not clear who within the longitudinal community is best placed to develop the shared narrative.

Transparency was also raised as a crucial component of any linkage communications strategy. Some delegates argued that to help make the case, longitudinal population studies needed to be more transparent on a number of levels, including the complex processes involved when undertaking data linkage, as well as what consents cover and how they were obtained. Clarity on these types of issues could help to reduce fear of the unknown, as well as misconceptions of the process, for example that data might be sold to other organisations and agencies. Ensuring participants are given regular feedback on how linked data are being used was seen as another important aspect of transparency.

While the buy-in of data holders and participants is obviously critical to the success of linkage projects, delegates also felt it was important to engage the general public on the issue, for example through community events. Public support for data sharing and linkage was considered a crucial element in persuading policy makers of the virtues of these processes.

Emerging opportunities for new forms of linkages

While many discussions of data linkage focus on administrative records, delegates also discussed opportunities to link to different forms of data, including spatial or social media data. Speakers shared experiences of working with such data, and the opportunities that open up when it is paired with rich longitudinal survey data. For instance, insightful studies of human behaviour and mental health emerge when information on social media use can be analysed in conjunction with the gold-standard measures of health and behaviour, alongside other experiences and characteristics, collected by longitudinal population studies. In another example,

linking longitudinal survey data to geographical information about participants' neighbourhoods allows for deeper understanding of how where people live affects their outcomes, over and above other characteristics.

However, delegates noted many challenges with these forms of data. Concerns about data quality were raised again, as well as issues around the utility of certain data for research purposes. For example, classifications and changes in geography over time can cause problems for analyses, and postcodes may not be as exact a measure as individual addresses (particularly in rural areas). Delegates noted the need for users to be adequately trained in using such data. Participant and public acceptability may be even more of a challenge with new or non-administrative forms of data for linkage, with greater concerns that social media linkage or GPS linkage could be a seen as too much like surveillance. However, participants may be more inclined to share such data if they see it as important to the wider social and scientific purpose of the study.

Including minority and vulnerable subgroups in linkage strategies

Certain subgroups, for example children in care, are difficult to retain in longitudinal population studies. However, they may have extensive data collected about them through other means, for example administrative records. Linkage may offer a viable alternative to studying such groups and a way of maintaining information on them in longitudinal studies. Delegates underlined that in order to understand the value of linkage strategies for subgroups – and crucially to gain permission from data holders – the longitudinal community needs to first assess current coverage, as well as the positive and negative characteristics, of different data sources. Building relationships with a range of agencies, such as local health providers and community groups, was considered important, particularly in the case of regional studies.

Resource and training needs

Discussions highlighted that data users are not adequately trained to analyse linked datasets. Upskilling users to take advantage of the opportunities afforded by linked

data would ensure the investment of time and money in ensuring these resources have strong returns.

It was also noted that effective communications with difference stakeholder groups would need to be properly resourced with qualified professionals.

New forms of data collection

Session lead: Lisa Calderwood, UCL Centre for Longitudinal Studies
Report author: Jon Tebbett, CLOSER

Key learning

- There is a need for an effective mechanism to share information on successes, failures, risks and benefits. This should include engaging with not just the longitudinal community but new disciplines, commercial companies, market research experts, funders and study participants.

- With many new forms of data collection, response rates can be low, particularly among certain groups. More feasibility studies are needed to understand what works well, including with specific demographics, and to realise the opportunities new methods can offer.

- The development of new technology demands skills that are not traditionally acquired in longitudinal study teams. Studies should work together and share resources to help reduce the heavy financial and resource burden.

- The use of new forms of data collection should be driven by science, and by utilising the best tool to answer a specific question. It should not be driven by technology and innovation for innovations sake.

Introduction

Longitudinal studies in the UK and internationally are increasingly collecting new forms of data. This reflects the opportunities in using digital technology to collect data that has been facilitated by technological innovation, the digitisation of everyday lives and widespread smartphone ownership. This technological innovation, and the resulting increase in forms of data collection, can offer improved granularity of information, immediate 'live' data collection, increased objectivity, and opportunities to collect new, novel measures. However, this brings

challenges for studies in terms of measurement, feasibility and whether these new forms of data collection are taken up by representative samples.

In terms of challenges, new forms of data collection can be considered within the context of the total survey error framework, looking at errors of measurement and errors of representation. Measurement error can be introduced through the different technology used by participants, for example if studies rely on participants using their own devices to take part. New technology can create barriers to participation, thus reducing representativeness, and highlights the importance of investigating how participation can be maximised.

Despite this, there are opportunities to make data collection faster, easier and cheaper. Technology can engage audiences and participants in new ways and reach audiences that traditional methods may not engage, and ultimately provide new ways for data to solve the complex issues facing society.

The session chair put the following questions to delegates for discussion:

- What are the most scientifically and strategically important challenges?
- What are the priorities for addressing these challenges?
- What should the priorities be for collaboration across longitudinal studies and how can this be achieved?

Summary of themes arising from discussions

The need to share knowledge and learning

As more studies experiment with new forms of data collection there is an even greater demand to share knowledge and help one another understand what works and what doesn't. Delegates wanted a forum to share successes, failures, and the risks and benefits of the methods they were using. It was emphasised that any new technology needs to be tried, tested and further developed, and that this could be done much more effectively together as a community rather than working in silos.

It was highlighted that the sharing of knowledge should take place beyond the longitudinal community. It was agreed that events like the CLOSER conference

were essential in making well-informed decisions, but there should also be a greater push to engage beyond this community. There would be benefits in collaborating with market research companies, commercial companies (e.g. technology providers, research agencies) and academics across all disciplines. Communities outside of longitudinal studies have expertise that could be utilised and new networks could be forged.

Delegates also expressed an interest in developing a mechanism to share feedback openly with funders about what works and the associated costs. There is a significant financial cost when using new technology for the first time, and this should be articulated to the funders. There was a demand for funders to fund more feasibility studies, and for the outcomes of such studies to be openly shared with the community. Others suggested that funders could consider grant awards to investigate specific methodology, for example the development of apps.

Delegates were keen to further engage study participants and involve them in decisions about the design of the data collection, especially in relation to new data collection methods.

How to evaluate data quality

The development of new technology is a fast-paced, changing environment where it can be difficult to identify the best methodology and to understand what is valid. This highlights the need for validation studies. However, delegates suggested that even when rigorous validation studies are undertaken they could rapidly become out of date as new, preferred, methods are developed. It is therefore difficult for studies to decide what technology or method to adopt and to identify the most important metric to record.

Measurement bias is a significant risk with new forms of data collection. Delegates identified the need to develop benchmarks for success, so the community can understand what a successful data collection would look like. This would help to highlight the best approach to measurement in terms of what is feasible and realistic. It could also help the community adjust for any bias introduced.

The nature of longitudinal studies, where repeated measurements are key, brings challenges as new technology continues to evolve. Technologies develop and change, and in the technology sector the changes are often very frequent which may result in differences in data collected over time within studies. This brings into questions whether repeated measures on apps are as useful as repeated measures in less frequent more traditional methods (such as questionnaires).

Challenges and opportunities around representativeness

Delegates highlighted how new technology can be effective at engaging with study participants in a way that is more meaningful to them. For example, young adults may engage with data collection driven by modern technology and older cohorts may prefer new passive collection methods. It is also the case that participants increasingly expect surveys to incorporate new technologies. However, moving away from traditional face-to-face methods might lead to a decreased sense in involvement in the study, and the rapid changes in method could come to be a burden on participants.

With many new forms of data collection, response rates can be low among certain groups. There are difficulties in terms of uptake, and studies can lose participants who do not have access to, or do not know how to use, technology. This introduction of sample bias was flagged as a key issue facing studies when adopting novel forms of data collection.

Delegates discussed whether the method of data collection could be tailored to the population under study. Different methods and technology might be more appropriate for use with teenagers compared with older people. This highlighted the need for more feasibility studies to understand what works well with specific groups. For example, there is an opportunity to understand the demographic characteristics of those most likely to download and use apps. Delegates also highlighted that although data collection through new technology does exclude certain groups of people, it can be effective at engaging others. The enthusiasm and increased use from such groups should be encouraged, engaged with and utilised by studies.

Scientifically driven, not driven by technology

The decision for incorporating new forms of technology should be driven by good science rather than innovation for innovations sake. New technology can seem attractive to funders and chime with a demand to be innovative. However, delegates promoted the need to prioritise the scientific drivers for research and the use of the best method to address that specific research query. For example, delegates discussed the need for good justification for developing an app. The cost associated with development is high and there needs to be a valid case for replacing an established method of data collection to answer the research question. There also needs to be consideration of the additional scientific value of a new approach through new information or improved quality of data.

Issues around technical skills and resources

The development of new technology requires skills that are not traditionally acquired in longitudinal studies. The acquisition of these new skills is expensive and difficult. The speed at which technology changes also means that the skills required often change with each new technology, again increasing the cost to the study. Delegates discussed potential models by which studies could access the skills needed. It was suggested that funding could be pooled, and technology could be shared. For example, experts could cover multiple studies or a single app template could be created that is shared across multiple studies. Delegates promoted the idea of investigating the sharing of apps, or app templates, across the community.

The delegates also discussed the challenges that can result from the introduction of third parties in the development of technology. There can be complicated legal issues, for example around privacy and data protection. In addition there is the possibility of the creation of 'black boxes' where it is difficult to ascertain exactly what information is being collected and how. In the case of apps, there are also high maintenance costs and this again is a difficult model to resource.

Resource and training needs

Delegates highlighted the need for an effective means of sharing information across the community. This should include engagement within the longitudinal community, but also reaching out to new disciplines, commercial companies, market research experts, funders and study participants. The community identified the need for a space to share experiences, learning and promote best practice.

Data discoverability

Session leads: Nic Timpson, Children of the 90s (ALSPAC), and Nathan Cunningham, Research IT and Innovation, University of Sheffield
Report author: Hayley Mills, CLOSER

Key learning

- All initiatives aimed at improving discoverability and interoperability must have clear objectives that are based on a strong understanding of the target users' requirements and research needs.
- There is demand for discovery platforms that provide highly detailed metadata, as well as for platforms that provide a higher level of information, and allow users to quickly find out if a certain study or data exist.
- Interoperability in discovery platforms can facilitate interdisciplinary research and reproducibility by facilitating research that combines data across domains, and ensure infrastructure adheres to FAIR meta(data) principles.
- Sustainable and interoperable infrastructure must be prioritised in funders' investment strategies, and the contribution of such work to novel research, open science and reproducibility recognised.
- There is a need for a discoverability leadership body with technical skills and knowledge, the ability to advise on standards, and a strong management team capable of coordinating across multiple stakeholder groups.

Introduction

The data and discovery landscape is complicated and varied with a portfolio of data resources for researchers and users spanning from: micro to macro data, conventional to specialist data types, study level to consortiums and commercial centres, general catalogues to federated searches and data pooling, as well as more

theme-focused platforms. All of these potentially have different aims, governance structures and access mechanisms. In this context, discoverability raises a number of challenges, including what we mean by the term, how we make use of existing resources, and what researchers' requirements are.

One aspect of realising the value of discoverability is the need for interoperability. There are several drivers for organisations at all levels to be more interoperable. Researchers and funders want to combine data from different disciplines to create new understanding and address policy objectives. Funders are also expecting (and in future will require) that data comply with FAIR (meta)data principles: **F**indable, **A**ccessible, **I**nteroperable, and **R**eusable. However, in many cases, projects and organisations are struggling to meet these standards as bespoke systems limit interoperability between platforms, which can have a knock-on effect on accessibility and reusability in particular. Standards are key to interoperability but the utilisation of standards across disciplines and organisational type varies massively, with most not using any standards.

The session chairs put the following questions to delegates for discussion:

- What are common challenges to discovery?
- What are challenges for semantic interoperability – how/what vocabularies, concepts and ontologies platforms are using & where do these lie?
- What are challenges for technical interoperability – use of metadata standards and the technical components relevant to sharing?
- What are the most scientific and strategically important issues?

Summary of themes arising from discussions

The purpose of discoverability

The general purpose of discoverability is to increase the use of existing data resources for secondary analyses, whether to answer scientific, policy and practice questions. Delegates had different interpretations about what discoverability meant more specifically, which was determined by several factors. The type of

data, discipline, audience (new or established), culture and historical context were discussed as factors affecting user requirements as well as expectation of discoverability. Establishing the target audience (user group) and their objectives can help inform how discoverability is understood.

Linked closely to understanding the purpose of discoverability, delegates discussed where in the research process discovery fits. For example, should discovery take place before any data are viewed? Should discovery platforms allow the user to make a decision about the data's utility? Should they allow the researcher to use and analyse the data within the same platform?

A continuum of discoverability was identified, from high- to low-level metadata, with demand for such detail differing across user groups and stages of the research process. Delegates agreed that both ends of the spectrum are valid forms of discoverability with different objectives and audiences. There is value in finding whether a study or data themselves exist, as well as demand for highly-curated metadata and harmonisation at a variable level, although providing these different levels of information for a diverse audience remains a challenge. Some delegates noted that many studies are not sufficiently resourced to produce detailed metadata, but even in the absence of funding it is important for studies to strive for some level of discoverability.

With the amount of data increasing and the many different types of data available, a way of knowing the property of data – who it belongs to and how to access it – is more important than ever. This is challenging, particularly for discovery platforms spanning multiple studies or data sources, in terms of establishing what processes to follow, setting users' expectations on the timescales for access, and balancing user needs with a multitude of data access arrangements. As there is not one definition or objective of discoverability, it is therefore likely not possible to have something that covers all levels, disciplines, data types and audiences, while meeting all user and organisational requirements. Delegates agreed that discoverability must therefore be purpose driven, but also deliverable across disciplines.

User expectations

Meeting user expectations is a challenge for any discoverability resource. If the aims of the intended users do not align with the aims of the resource, this may result in the user expectations not being met. In addition, there is not necessarily alignment between users, data collectors and metadata curators, for example in terminology and culture.

Differences in the way users search for information can also be challenging to accommodate within a single resource. Users will carry out searches differently because of different perspectives, or they may not know how to search effectively on a particular platform, or may not be familiar with the terminology the resource requires. These issues may be particularly challenging to address on platforms that offer only high-level metadata, compared to platforms where a more granular level of metadata are available to search, which allow greater search capacity and filtering.

Interoperability

At its simplest definition, interoperability is the ability for one thing to communicate to another, whether that is human-to-human, human-to-machine or machine-to-machine. There needs to be a common language to do this, and a common understanding of what it is to be interoperable. Interoperability and standards are important parts of any discovery strategy, to allow federated or other ways of aggregating discovery across multiple resources. However, the meaning of interoperability more specifically can be affected by access restrictions, attitudes, policies and geography.

Interoperability was discussed as being important for reproducibility, as well as maximising the benefits to funders of data reuse, including achieving more value from the existing resources and a greater return on investment. Interoperability can also help users better understand the data they are using, as schema or standards help with labelling and interpretation of what the data are.

Delegates noted that existing resources that have not aligned on any standards will face significant challenges in achieving interoperability with other platforms, and balancing interoperability with their existing functionalities. In order to determine what level of interoperability is required, delegates agreed that there is a need to set out clear objectives first, so that resources meet user requirements. These objectives will also inform the level and type of interoperability required. For instance, interoperability at the variable level may allow better support for harmonisation activities.

The current lack of interoperability is historical and differs across cultural domains. Many challenges were discussed relating to the different environments, including governance, complexity of data, data types, legislation and temporariness of projects. However, there are some basic models in social science that demonstrate technical and semantic interoperability resulting in more consistent and meaningful discoverability. There are barriers to achieving such models, but using standards for structured (meta)data can help technical interoperability, as substantial portions of many existing standards are interoperable. Semantic interoperability may be more difficult to achieve, as the level of mapping required between different vocabularies is often missing. Furthermore, although there are many vocabularies and standards available, the adoption of these is limited, and when they are adopted, they may not be used consistently.

Delegates recognised a role for funders in pushing interoperability, namely by investing in long-term infrastructure and projects that use existing standards.

Prioritising investment in infrastructure

Upgrading infrastructures, moving to new technologies and general day-to-day (meta)data management is often undervalued and under-resourced, which results in missed opportunities in metadata reuse and efficiency savings. The importance of such infrastructure in making possible cutting-edge research, open science and reproducibility needs to be recognised by funders and reflected in their investment strategies.

Delegates identified that funding is needed within studies in order to improve infrastructures and data management. They also recognised that study teams can struggle to balance the investment of time required for managing a data resource with the requirement to carry out their own research for which they have been funded. However, it was broadly agreed that even with little or no funding, it is still necessary for studies to consider strategies for discoverability and interoperability.

Finally, delegates felt there was a need for long-term funding plans to ensure researchers have access to well-run, sustainable resources for data discovery. There also needs to be recognition that highly-curated metadata is more expensive to achieve, particularly if this is not available for legacy data, so systems for producing metadata sustainably are required.

Leadership

Delegates identified the need for leadership in data discoverability, particularly in areas such as interoperability and advising on standards, where coordination is required across numerous stakeholder groups. However, delegates were not clear who should be taking the lead and whether this should come from funder or researchers, although it was noted that researchers receive little recognition for their work on data or infrastructure outputs. A leadership body would require technical skills and knowledge, as well as a good management team.

Resource and training needs

In addition to resources outlined above, delegates expressed a need for best practice guidance and training on what factors to consider in the design of discovery projects, particularly in terms of discoverability, interoperability and sustainability. There was also interest in training and resources on the importance of metadata and using standards, including a 'train the trainers' model that could allow research data managers to spread knowledge and understanding more easily through larger institutes.

Reducing attrition and participant engagement

Session leads: Hazel Inskip, Southampton Women's Survey and Andy Wong, MRC
National Survey of Health and Development
Report author: Meghan Rainsberry, CLOSER

Key learning

- While monetary incentives are not universally appropriate or affordable, most studies make use of some form of 'reward' for participation.

- Participants' family members remain critical influences on participation throughout the life course, and studies should treat them as allies in participant engagement strategies.

- Cultural and country context deeply affects tracing and participant engagement efforts, both in terms of the barriers presented, and the appropriateness and practical viability of different solutions.

- Survey design can be a key engagement tool – making participation flexible and fun can raise response rates.

- When engaging participants, most studies are keen to move towards a model of tailored communication for different sub-groups, but not all can afford it.

- To improve participant engagement, the longitudinal community needs adequate financial resources and a strong literature base on what works.

Introduction

All longitudinal population studies lose participants over time. The unavoidable reasons are mortality, and for some studies, emigration. But other participants are lost to studies for other, modifiable reasons, such as non-response or withdrawal. Attrition leads to a loss of statistical power, loss of representativeness, and thus

scientific value of the data. As such, most studies take proactive steps to engage participants in order to maintain sample size over time.

However, many studies are struggling with retention of participants in the modern age, with social changes affecting willingness to take part, such as migration, family breakdown, privacy concerns, apathy and dissolution of trust, populism and increasingly busy lifestyles. Methods for engaging participants are equally varied. Newsletters, postal mailings and other forms of one-way bulk communications fall at one end of the spectrum, with participant advisory groups and other forms of direct involvement at the other. Studies are experimenting with incentives, new technology, social media and other means to keep participants engaged over time.

The session chairs asked delegates to consider:

- What are the main challenges in reducing attrition and participant engagement?
- If you could do anything, what would you want to do to reduce attrition and engage participants?
- What are the barriers?

Summary of themes arising from discussions

Incentives

Much of the discussions centred around incentives, however attitudes and approaches differed significantly across cultures. In more transactional cultures, (often large) monetary incentives were part and parcel of administering the surveys. However, participants in other cultures were averse to financial rewards for taking part.

A huge range of non-monetary incentives are being used by studies across cultural and country contexts. The main categories included prizes, giveaways or small gifts, as well as social interactions, like reunions or virtual communities. Cards (namely birthday and thank you cards) were very popular among participants. A few studies mentioned the role of media coverage in encouraging participants to

take part – seeing the study in the news is a major motivation for many participants.

One major 'reward' for taking part in some studies was the return of individual level health data – with participants seeing the surveys as an opportunity for a 'check-up'. However, there were significant reservations among delegates that this form of feedback could change behaviour and influence the future life paths of participants.

Finally, many studies are required to replenish their samples periodically by recruiting new participants. New recruits were not seen as having the same sort of motivations or study loyalties of existing members. In these cases, incentives were seen as more critical to securing participation.

Family members as key allies

Many delegates emphasised the important role that family members play in keeping participants in a study – throughout the course of their lives.

In the early years, parents or guardians act as gatekeepers, as many studies need to go through the main carer to invite child participants to take part. However even as participants grow older, their families continue to provide encouragement to stick with the studies. Delegates spoke of sibling pairs nudging each other to stay involved, and of 'family loyalty' being a top reason for panel participants to keep up their participation.

Tracing

Keeping track of participants' locations in order to maintain contact with them is a fundamental to the operation of longitudinal population studies. However, several delegates noted that their tracing efforts can be stymied by lack of funding, ethical concerns and infrastructure. Failure to carry out tracing activity between waves, or in the more immediate lead up to fieldwork, can result in costly efforts to find participants once a survey is underway.

While low- and middle-income countries face unique challenges, such as lack of reliable communications technology or travel infrastructure to reach participants in remote locations, it was equally interesting how many challenges they have in common with studies in high income countries. Coping with migration of participants, family breakdown and gaining trust are all barriers shared across country contexts. On the other side, community-connectedness in many developing countries can boost tracing efforts – but this is something that is on the decline in Western cultures.

As noted above, family members were key partners for most studies in tracing participants and securing their participation, but delegates also mentioned the importance of schools, GPs, and secondary contacts for tracing.

Participant burden

While all delegates agreed that participant burden was a major factor in participation, what constitutes a 'burden' seemed to be in part culturally driven. For example, some suggested participants in Western cultures may look favourably on a shortened or online questionnaire, whereas other cultures may see shorter visits or lack of personal contact as rude or abrupt.

However, there was consensus that offering flexibility, such as short and long versions of questionnaires, as well as 'fun' data collection methods could help engagement. Also, modernising certain elements – such as online booking systems for interviews – has been used successfully to boost response.

Messaging and tailoring of information

Getting across the value of the studies was seen as a central and shared challenge. Many delegates urged greater collaboration on articulating the unique contribution of longitudinal evidence, and the impact of the studies on wider society. However, others noted that participants can be disappointed by the perceived lack of political progress on social and health problems highlighted by the studies, when evidence is not acted upon.

Many studies struggled with similar challenges in articulating the data collection and linkage processes, and data security in a clear, simple and engaging way. Many attempts have been made to use infographics and other visuals to depict this information, but few have tested the effectiveness of these methods.

Finally, many delegates discussed the need to tailor participant communications to different demographics and motivations. For example, men and boys are known to be at greater risk of dropping out of studies, and developing communications to engage them more directly was seen as important. Delegates felt that in an age when information can be easily and extensively personalised, it was 'old-fashioned' not to tailor participant communications. However, once again cost and insufficient evidence of what works were seen as barriers to tailoring content.

Resource and training needs

When asked about barriers to employing participant engagement methods, two main issues were raised: cost and lack of evidence.

For many studies, participant engagement strategies are driven by budget limitations, and few have adequate resources to evaluate efforts or design strategies based on available evidence. Delegates expressed a strong desire to promote and develop the literature on what works, in what context, and for which populations. In particular, studies wanted evidence on critical ages for attrition and how to best manage transition points in participants' lives.

There was some suggestion that studies need to emphasise to funders that certain participant engagement costs are a central part of the science of the studies, rather than a 'nice-to-have'.

Appendix: Conference programme

09:30   Registration and refreshments

10:00   **Welcome to the conference**
Prof Bobby Duffy, CLOSER Advisory Committee Chair

10:05   **Panel discussion - Setting the scene: our greatest challenge**
Chair: Prof Rebecca Hardy, CLOSER Director
Panellists: Alissa Goodman, Director of UCL Centre for Longitudinal Studies; Susan Morton, Principal Investigator of Growing Up in New Zealand; Shane Norris, Principal Investigator of the Soweto First 1000 Days Cohort & Co-Principal Investigator of the Birth to Twenty Plus cohort; Narayan Sastry, Associate Director of the Panel Study of Income Dynamics; Nic Timpson, Principal Investigator of the Avon Longitudinal Study of Parents and Children

11:05   Short refreshment break

11:30   **Parallel sessions A**

A1 Reducing attrition and participant engagement: international perspectives	A2 Data linkage	A3 Data harmonisation
Session lead: Hazel Inskip (Southampton Women's Survey) Speakers: Kate McGonagle (Panel Study of Income Dynamics), Johanna Choumert-Nkolo and Luca Privinzano (Tanzanian Kagera Health and Development Survey)	Session lead: Sally Bridges (Born in Bradford) Speakers: Galina Daraganova (Australian Institute of Family Studies), Oliver Davis (University of Bristol)	Session lead: Dara O'Neill (CLOSER) Speakers: Eoin McElroy (University of Leicester), Sarah Bauermeister (Dementias Platform UK)

12:30   Lunch and networking

13:45   **Parallel sessions B**

B1 New forms of data collection	B2 Data linkage	B3 Data discoverability in the biomedical sciences
Session lead: Lisa Calderwood (Next Steps) Speakers: Jon Burton (Understanding Society), Melissa Wake (Generation Victoria)	Session lead: Andy Boyd (ALSPAC) Speakers: Emla Fitzsimons (Millennium Cohort Study), Oliver Duke-Williams (CeLSIUS)	Session lead: Nic Timpson (ALSPAC) Speakers: Barry Radler (Midlife in the United States), Phil Quinlan (HDR UK Atlas Project)

14:45   Short refreshment break

15:00   **Parallel sessions C**

C1 Data harmonisation	C2 Reducing attrition and participant engagement	C3 Data discoverability: issues in interoperability
Session lead: Rebecca Hardy (CLOSER) Speakers: Angela Pinot de Moira (LifeCycle), Susan Morton (Growing Up in New Zealand)	Session lead: Andrew Wong (MRC 1946 National Survey of Health and Development) Speakers: Christian Beron and Nathalia Martinez (Uruguayan Nutrition, Child Development and Health Survey), Paz Garcia (Twins UK)	Session lead: Nathan Cunningham (University of Sheffield) Speakers: Jon Johnson (CLOSER), Mari Kleemola (Finnish Social Science Data Archive and CESSDA)

16:00   **Reflections from the day**