House of Commons Science and Technology Committee 'The right to privacy: digital data' inquiry

Response from CLOSER, the home of longitudinal research (UCL Social Research Institute)

Authors: Rob Davies, Head of Policy and Dialogue; Jon Johnson, Technical Lead; Dr Neil Kaye, UCL Research Fellow; Hayley Mills, Metadata Manager
Reviewers: Professor Rebecca Hardy, Director; Jon Tebbett, Project Manager

1. About us:

1.1 CLOSER[i], the home of longitudinal research, is the interdisciplinary partnership of leading social and biomedical longitudinal population studies, the UK Data Service and The British Library. Our vision is to increase the visibility, use and impact of longitudinal population studies, data and research to ensure that longitudinal evidence is used to address the health, social, economic and environmental challenges facing the UK, now and in the future.

1.2 Our studies[ii] comprise of both national and regional longitudinal population studies from across the UK. They include the British Birth Cohort Studies, ONS Longitudinal Study, English Longitudinal Study of Ageing, Born in Bradford, Southampton Women's Survey, Avon Longitudinal Study of Parents and Children, Generation Scotland, Understanding Society – the UK Household Longitudinal Study, and more.

1.3 CLOSER has been funded by the UKRI Economic and Social Research Council (ESRC) since 2012 and is based at the UCL Social Research Institute.

1.4 The ESRC's suite of longitudinal population studies provide an unparalleled shared research resource. By collecting data at intervals from the same people over their lives, these studies are continuously building an ever-richer data resource that captures how life in the UK changes over time and how events and exposures have long-term impacts.

2. Our reason for submitting evidence:

2.1 CLOSER is a significant ESRC data infrastructure investment. Our pioneering work over the past 10 years has helped to improve the transparency and availability of high-quality data from longitudinal population studies and our contribution to metadata standards and documentation of harmonisation processes will prove vital for ensuring data are reusable and reproducible.

2.2 Our flagship resource, CLOSER Discovery[iii], is the UK's most detailed search engine for longitudinal data. It is a free online platform that enables researchers and analysts to search through the rich metadata from UK longitudinal population studies and see, at a glance, which studies and data meet their research requirements. It is a user-friendly resource for locating the variables that best suit an individual's research interests, and testing their robustness. The platform is built using metadata. By using metadata, CLOSER Discovery enables users to find out what is available in the UK's longitudinal datasets, understand the context of how the data were created, and make an assessment about whether the data is relevant for research.

2.3 CLOSER brings together representatives from the social and biomedical longitudinal data and research community, government analysts and policy makers, and research funders to discuss key challenges and identify solutions, including through our Communities of Practice and *Preparing for the Future* conferences.

2.4 CLOSER manages several Communities of Practice[iv], including a Data Managers Network and Data Linkage Working Group. These networks are designed to cultivate knowledge exchange, share best-practice, and facilitate innovative and collaborative projects.

2.5 Other areas of our work related to this inquiry include the development of software and tools for data management, training and capacity building, improving interoperability between data infrastructures, and developing the use of new technologies (such as machine learning in social sciences) to gain new insights into existing data.

2.6 Our response focuses on the following questions in the call for evidence:

- The potential benefits, including to research, to effectively use and share data between and across Government, other public bodies, research institutions and commercial organisations, and the existing barriers to such data sharing.
- The extent to which data issues are appropriately addressed by the Government's National Data Strategy, its draft strategy, data saves lives: reshaping health and social care with data, and its consultation Data: a new direction.
- The extent to which appropriate safeguards and privacy are applied in the usage and sharing of individuals' data.

3. Key messages:

- Considerable value emerges when we link the rich and extensive information collected about individuals through their participation in social and biomedical longitudinal population surveys with the detail that comes from administrative data held about the same people.
- There are a number of barriers to accessing data held by government, including legislative uncertainty, resource (including insufficient staff with the necessary data management skills and high staff turnover), out of date data management structures or procedures, and complications and delays with data access requests.
- Setting appropriate and well-defined standards is a crucial first step in unlocking the value of data.
- The provision and consistent use of metadata to an established standard is fundamental for efficient and meaningful exchange (sharing) of data between organisations.
- There is a need for a continued focus by the Government on data quality, common standards, use of metadata, and data skills. These issues should be treated holistically, not as individual, disconnected initiatives.
- Fundamentally, data needs to be 1) discoverable, 2) useful and 3) usable.

4. **The potential benefits, including to research, to effectively use and share data between and across Government, other public bodies, research institutions and commercial organisations, and the existing barriers to such data sharing:**

4.1 The UK is home to a remarkable set of scientific studies that have tracked generations of people growing up in Britain over the last 70 years. These longitudinal population studies are unique in science and unparalleled elsewhere in the world – no other country has anything like them on the same scale. Findings from these studies have shaped the world we live in today, changing policies around birth, schooling, social mobility, parenting, ageing and more. More recently they have helped to understand the health, social, economic and behavioural impacts of the COVID-19 pandemic at both a national and regional level, and across all generations and ages[v]. Often referred to as the "jewel in the crown of UK science" they have touched the lives of almost everyone in Britain today, however data and evidence from these studies are still underutilised.

4.2 Government departments routinely collect data on various aspects of life in the UK: children's progress through the education system, information about benefits claimed and taxes paid, and individuals' experiences of hospital treatment. Researchers are interested in this 'administrative data' because its volume and detail can vastly exceed what it's possible to collect through other routes such as surveys. It is widely recognised that these data have immense potential value for research across a wide range of subject areas.

4.3 Considerable value emerges when we link the rich and extensive information collected about individuals through their participation in social and biomedical longitudinal population surveys with the detail that comes from administrative data held about the same people[vi]. This kind of linkage, done with the consent of the person concerned, overcomes one of the major problems with administrative data; it isn't collected with research in mind and so will never collect information about everything of relevance to a particular research question. Survey data isn't perfect either; it rarely collects information that has the enviable detail and frequency found in an administrative dataset, and also requires input and effort from survey participants. So though both types of data are important and have considerable value, the picture gained by combining them is far greater than that which emerges when each is analysed on its own; for 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.

4.4 There are many complexities and challenges associated with linking survey and administrative records. In 2018, 2020, and 2022, CLOSER hosted its *Preparing for the future* collaborative conferences[vii]. Delegates from across the social and biomedical sciences came together to discuss the issues facing longitudinal population studies now and in the future (including data discoverability, data harmonisation, and data linkage), share best practice, and identify ways to tackle key challenges. Unlocking opportunities for enhancing longitudinal population studies through linkage to administrative data, particularly through improving access to administrative data, has been a recurring theme throughout these three conferences.

4.5 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 *(data quality)*.

4.6 Although there have been successes in this area[viii] and new opportunities for linkage between different data sets are constantly emerging, common challenges persist around data access processes and procedures, data quality, research utility, and participant and public acceptability. Furthermore, the potential offered by survey and administrative data linkage are often hindered by cultural barriers within government departments and disparate and opaque data access processes and procedures. In our experience, there are a number of barriers to accessing data held by government, including legislative uncertainty, resource (including insufficient staff with the necessary data management skills and high staff turnover) and out of date data management structures or procedures. There are also often complications and lengthy delays with data access requests.

5. The extent to which data issues are appropriately addressed by the Government's National Data Strategy, its draft strategy, data saves lives: reshaping health and social care with data, and its consultation Data: a new direction:

5.1 There is a need for a continued focus by the Government on data quality, common standards, use of metadata, and data skills. These issues should be treated holistically, not as individual, disconnected initiatives.

5.2 Historically, there have been major **issues with the quality of data** in some government-held data sets, including inconsistent collection of data, 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). The emergence of quality statements is a helpful way of understanding the quality of available data, however there needs to be a review of its effectiveness in improving data quality over the long-term.

5.3 **Setting appropriate and well-defined standards is a crucial first step in unlocking the value of data.** The recent adoption by Office for National Statistics (ONS) of a set of data and metadata standards[ix] is welcome. However, **to roll this out across Government will require a program of training, advocacy and the development of common software across different departments, in order to turn this into genuinely shareable data.** Data providers and data consumers should be encouraged to utilise these standards. The benefits could be enormous, for instance, the internet hotel and flight industry would not exist were it not for the agreement to use the same data standards and metadata formats.

5.4 **To maximise the use of data across the UK, there is a need for comparable ontologies**, for example the European Language Social Science Thesaurus[x] uses the same terminology with translations so that you can locate the same information across different European archives – this approach needs to be adopted in the UK. The adoption of the recent suite of metadata standards[xi] could be transformational in not only providing the basis for efficient exchange of data, but also to combine data meaningfully. The recommendation by the ONS to use the Asset Description Metadata Schema (ADMS) would (if adopted widely) be a major step forward in allowing data to move seamlessly within government.

5.5 Agreement on the structure and format is not in of itself sufficient to achieve such an ambition. The European Commission's adoption of the ADMS recognised the need for a co-ordinating function for rationalising the terminologies and put in place a Metadata Registry to co-ordinate these terms and vocabularies. This has been instrumental in ensuring data quality and the reliability of data collection.

5.6 **What is metadata and why is it important in the data landscape?**

Metadata[xii] – data about data – is hugely important for effective data use as it contains crucial information necessary to exploit the full potential of datasets for research. Structured metadata defines the relationship between data items to enable computer systems to understand the contextual meaning of the data, for example, to display the relevant information on a website. Structured metadata tells a computer what something is, how it relates to other objects and what to do with it. By standardising the content and structure, it makes it easier for computers to automatically extract information from the metadata. This information can then be provided to researchers to help them discover and access data from many different sources. This is what makes data FAIR (Findable, Accessible, Interoperable, and Reproducible). Metadata represent a common mechanism for communication between researchers and must not be overlooked when considering data sharing strategies. **It facilitates data sharing and allows data collected in one study to be re-used in the future by other researchers.**

5.7 **Data skills are a fundamental pillar of effective data use and should remain a priority of the National Data Strategy and further work in this space.** This is an area we feel requires more focus: CLOSER's training needs review[xiii] identified several fundamental issues in the level of data skills across both research and policy domains. These include: lack of specific analytical skills and software literacy, inability to access or handle data and to physically access data (for example, those held in a secure space), a lack of training provision for specific techniques, and poor awareness of the complexity of longitudinal population data. **Increasing training provision and removing barriers to training access should be explored in further depth by the government to ensure that users are better equipped to analyse, interpret and understand a wide range of data, including longitudinal population data.**

6 The extent to which appropriate safeguards and privacy are applied in the usage and sharing of individuals' data:

6.1 The relationship with participants in the CLOSER longitudinal population studies span across a number of decades, presenting a unique opportunity for others to understand how people view their data being used for research purposes. The CLOSER studies invest considerable resource to understand public expectations and to maintain participant trust, for example, studies use public and participant feedback to steer the development of data security and safeguarding facilities that allow research on linked routine records while maintaining legal compliance and public acceptability. We would urge learning from the experience from these long-standing relationships and that the measures longitudinal population studies have in place should be considered when developing plans around trustworthiness of data collection, usage and how these are communicated to study participants and the wider public.

6.2 There are robust procedures for accessing the data from longitudinal population studies. The ESRC-funded studies are licensed through the UK Data Service (UKDS)[xiv] to bona fide researchers for not-for-profit use. Anonymised data are deposited for use by the research community at the UK Data Archive, based at the University of Essex, and available via the UKDS. For more potentially sensitive or disclosive, but still anonymous, information the UKDS operates an application-based secure lab facility for accredited UK Higher Education Researchers only.

6.3 All research studies funded by the Medical Research Council (MRC) must comply with the 'MRC Policy and Guidance on Sharing of Research Data from Population and Patient Studies'[xv]. This sets out the MRC's requirements and expectations for studies on matters including data standards, data sharing, the governance of data access, facilitation, and data-sharing agreements.

28 January 2022

[i] https://www.closer.ac.uk/

[ii] https://www.closer.ac.uk/timeline/

[iii] https://discovery.closer.ac.uk

[iv] https://www.closer.ac.uk/our-networks/

[v] https://www.closer.ac.uk/covid19-longitudinal-research-hub/

[vi] https://www.closer.ac.uk/research-fund-2/data-linkage/

[vii] https://www.closer.ac.uk/preparing-future/

[viii] https://www.gov.uk/government/publications/joined-up-data-in-government-the-future-of-data-linking-methods/joined-up-data-in-government-the-future-of-data-linkage-methods

[ix] https://www.ons.gov.uk/aboutus/transparencyandgovernance/datastrategy/datastandards

[x] https://elsst.cessda.eu

[xi] https://www.ons.gov.uk/aboutus/transparencyandgovernance/datastrategy/datastandards

[xii] https://learning.closer.ac.uk/learning-modules/understanding-metadata/

[xiii] https://www.closer.ac.uk/wp-content/uploads/CLOSER-Analytical-training-needs-review-full-report.pdf

[xiv] https://ukdataservice.ac.uk

[xv] https://www.ukri.org/publications/mrc-guidance-on-sharing-research-data-from-population-and-patient-studies/