ESRC Longitudinal Studies Review 2017

Further analysis of responses to the consultation

Paper 7:

Data linkage

April 2017

Report author:

Ruth Townsley, Independent Researcher

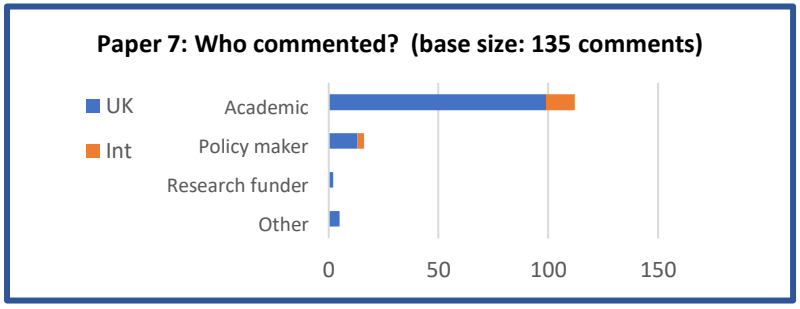
The views represented in this report are from those who responded to the consultation and do not represent the views of ESRC

Introduction

The ESRC Longitudinal Studies Review 2017 is exploring the current and future scientific and policy-relevant need for longitudinal research resources. The review started in October 2016 and will report to ESRC Council early in 2018. An open online consultation in autumn 2016 sought input broadly, resulting in 637 completed responses from UK (83.4%) and international (16.6%) respondents. Respondents were predominantly from the academic sector (80%) as well as government, civil society and business sectors (20%). The main findings of the consultation were published in December 2016 in an initial report. This report is supplemented by short briefing papers that examine key themes from the consultation data in more detail.

Briefing paper 7: data linkage

Data linkage is a process which brings together two or more sets of administrative or survey data to produce information which can be used for research and statistical purposes. The issue of data linkage was the most frequently mentioned methodological sub-theme with 120 coded comments in answer to survey question 9. A search of the whole dataset identified a further 15 comments relating to data linkage (in answers to questions 8 and 10).



Around 40% of these comments simply stated a need for data linkage, with little additional input offered. The remaining 60% of comments were more detailed and provided a rich source of material on current thinking across the longitudinal studies community.

Why do respondents think data linkage is important?

Respondents talked about the potential of administrative data linkage to add value to publicly funded research by enhancing (but not replacing) the research capacity of new and existing longitudinal studies, both scientifically and methodologically.

Respondents suggested that methodologically, data linkage could:

- Add detail and frequency rarely possible in surveys, thus improving data quality
- Help to fill gaps, both between waves and within existing datasets, e.g. by linking to social media accounts, health record, tax records, education outcome data, and records on income and participation in government programs
- Reduce the input and effort needed from participants, potentially reducing attrition
- Lead to fewer questions, hence less time and resource needed for data collection and analysis
- Help to significantly save money if existing health data from NHS records could replace the collection of new biodata
- Provide unique opportunities to develop new methodologies

"Data Linkage is already being achieved in some longitudinal studies but it offers considerably more potential. Unfortunately, progress with data controllers has been slow. As this data becomes more routinely accessible, it will be important to better understand the relative value of how survey and administrative evidence best complement each other and the potential for administrative data to 'fill gaps' between data collection waves." (ID 17)

Scientifically, respondents suggested that linkage between longitudinal resources and other non-research data could:

- Help to derive nationally representative samples
- Be useful for contextualising research findings
- Help to answer a vast range of new research questions such as: effects and impact of contact with state agencies; links between educational outcomes and many other areas of life such as income, home ownership, criminal activity, etc; influence of social networks, including effects of social media, on longer term outcomes; effects of pollution, car use and other areas of local-area data
- Have the potential to be two-way, e.g. longitudinal studies can also add detail and context to administrative datasets.

"Linkage of longitudinal data to geographical information systems represents a substantial challenge and opportunity with a huge amount of potential off-shoots if this data could be refined into a highly useable format. For instance, the Add Health study in the US provides indicators such as the density of alcohol outlets in the census-tract level area as well as 'obesogenic' features of the environment (proximity to parks, street connectivity, crime levels) that are readily available to researchers and have been used to good effect in many studies." (ID 312)

"Next Steps has already demonstrated the potential and value of linkage to educational records by linking to the National Pupil Database. The value of this linkage is demonstrated by an increasing volume of research which is utilising the linked data, not only to ascertain the predictors of young people's educational outcomes, but also to explore how test scores relate to a young person's family context and to teachers' rating of students. By seeking additional data linkages for Next Steps such as health records, economic records, HM Revenues and Customs, UCAS information, ILR and HESA data will provide an invaluable opportunity to explore methodological aspects such as survey response and missing data as well as enhance our substantive understanding of theses domains in these young adults' lives." (ID 446)

Being aware of the benefits and challenges of data linkage

Many respondents made comments that highlighted the potential benefits of data linkage alongside the concurrent challenges. It was clear that for activity in this area to be successful, the possibilities and implications of data linkage need to be recognised and understood. For respondents, these included:

- An awareness of what data sources exist, their coverage and prioritisation by the devolved administrations, and how the datasets could enhance longitudinal resources
- Understanding the strengths and challenges of different data sources in terms of data coverage, completeness, quality and accessibility, e.g. in England, respondents felt that the National Pupil Database and HMRC records have good coverage and completeness, but other datasets (e.g. Police National Computer Records, Hospital Episodes Statistics and data from HSCIC) are more difficult to assess and analyse, and the level of missing data is not well understood
- The potential impacts of data linkage on the operation and management of some studies in the longer-term – e.g. data linkage can save money and time but may mean that field-work staff may need to be re-deployed or trained to collect different/new data, or that resources could be used elsewhere

A few people gave examples of successful data linkage activities which could potentially be used as case studies to demonstrate possibilities and inform learning (e.g. ID 446). It will be important to continue to map and document processes and outcomes of efforts to link administrative and longitudinal research data so that learning can be shared effectively.

"We negotiate data sharing with the DfE, SFA and HESA. This has been very time consuming and it has not always been easy to establish a contact within these agencies. We would like more information on the range of government data available and for it to be more easily accessible to researchers." (ID 343)

Access, infrastructure and ethical issues

Respondents raised a number of practical and logistical issues regarding access, infrastructure and ethical issues relating to data linkage, as summarised below:

- Access to administrative data – the need for this to be easier and quicker was very frequently mentioned:
  - Respondents were unclear why some datasets were reasonably straightforward to access (e.g. NPD) whilst others were more challenging (e.g. HES) or were simply not available at all (much DWP data)
  - One person was using an international dataset because of being unable to access the UK data required
  - It took one study seven years to be granted access to an essential government dataset
  - A few people mentioned that CLOSER, CLS, some studies and the research councils are working to facilitate easier and more efficient access through direct work with government departments and through the Digital Economy Bill, and this was welcomed

- Infrastructure for data linkage – the need for strategic work to create a national infrastructure was mentioned by some respondents and comments included:
  - The need for a common approach to working with data controllers and negotiating data sharing agreements with government departments to save time and avoid duplication
  - Agreement on best practice in collection/retention of consents and sharing and dissemination of linked data with a secure environment
  - Potential for some linked files to be made available under special license with documentation on linkage methodology
  - The need for dedicated resources for the lengthy and complex work involved in (a) assessing the quality and utility of potential datasets; (b) developing statistical procedures to link them; (c) cleaning them; (d) developing methodologies for further analysis; (e) developing documentation and training
  - A few respondents mentioned the ESRC and MRC funded infrastructure work being conducted by UKDS, ADRN, CLS and the Farr Institute

- Ethics and governance – many people had things to say about the ethical and legal challenges associated with data linkage:
  - Consent – obtaining agreement for unconsented use of linked data – how can consent be maximised?
  - One respondent mentioned consent exceptions under Section 251 of the NHS Act 2005
  - How to ensure anonymisation and security in the context of linked data from a wide range of data sources
  - The legal framework is generally restrictive rather than permissive, despite the fact that there is some evidence to show that existing cohort members mostly give consent to data linkage – how can this process be streamlined?
  - A few people felt there was a risk averse culture amongst data controllers even when use of data is legally permissible
  - UKDS has established a framework for data linkage which includes ISO security measures – this has led to improved data linkage with several government departments.

"Although access to education records (such as the National Pupil Database) has to date been relatively straightforward, linkage to other administrative data of relevance to education (for example, relating to health or social background) remains far more challenging. This impedes the potential use of administrative data for education research." (ID 52)

"[We have] developed a highly successful model of onward sharing of administrative records together with DfE via the UK Data Service (UKDS). We are now in discussion with a large number of other data controllers (including HSCIC, HMRC, DWP, MOJ) to extend this model of working further. While to date other Government departments have been reluctant to allow similar data dissemination via UKDS, we are now optimistic that more will start to accept this model given the ideal framework offered by UKDS in terms of its security arrangements (ISO27001) and good governance e.g. the Five Safes principle". (ID 689)