Institute of Education

UCL

Embedding Centre for Longitudinal Studies cohorts into national linked administrative datasets – review

By Karen Dennison

CENTRE FOR LONGITUDINAL STUDIES

UKRI Economic and Social Research Council

Summary

There is strategic interest in embedding cohort data into linked administrative datasets and CLS should continue working with ONS, ADRUK and relevant government departments and organisations across the UK, such as SAIL in Wales, to enable this to happen.

- Legally there should be no barriers for cohort data linkages into linked national administrative datasets.
- From an ethics and consent point of view there should also be no major barriers to linkages for consents already obtained. However, it would be challenging to link to linked population datasets for which we do not have consent to all elements and this needs further exploration.
- There are also no major technical barriers to the data linkages
- CLS should feed into the ESRC-HDRUK Admin Spine Scoping Study
- A body of work is needed around consent bias to be clearly documented for potential researchers.
- CLS should put together scientific use cases for data controllers around the potential use of cohort data embedded into linked administrative datasets.

Contents

Introduction

In this project we undertook a short scoping review related to a number of strategic aspects of our record linkages programme. This report focuses on opportunities for embedding CLS cohorts into national linked administrative datasets.

Background

CLS has a programme of data linkage for the four cohort surveys that it runs and which is based on explicit informed consent from cohort members (CMs) to link health, education, economic and crime records into their survey data and to make the linked data available in de-identified form to researchers.

To date, this programme has focused on data linkages where administrative records are linked into the cohort data and the resulting datasets are deposited with the UK Data Service and made available under secure access arrangements. These linkages have so far been limited to single administrative datasets e.g. we have linked Millennium Cohort Study (MCS) data to health records and also separately to education records.

This review looks at the possibility of linking consented cohort data into national linked administrative datasets, meaning that researchers would have access to population level data for a range of administrative data alongside a set of richer information for a subset of that population.

What we did

We spoke with the Office for National Statistics (ONS), including staff from its leadership team and staff with expertise in the areas of longitudinal, health and administrative data. We also spoke with the Department for Education (DfE) and the Department for Work and Pensions (DWP) in the context of the Longitudinal Educational Outcomes dataset. We also spoke to the Co-Directors of SAIL based at the University of Swansea, and the lead for the Economic and Social Research (ESRC) Council and Health Data Research UK (HDRUK) Administrative Spine Scoping Study. Our aims were to find out what linked administrative datasets are, or will, be available, if there is willingness to embed the CLS cohorts into national linked administrative datasets, any legal, ethical or other challenges around this and how we would go about doing it and making the data available for research purposes.

National linked administrative datasets and interest in embedding cohort data

There are a number of potential administrative datasets and programmes of work of interest –

National linked administrative datasets and interest in embedding cohort data

There are a number of potential administrative datasets and programmes of work of interest. These include -

The Longitudinal Educational Outcomes (LEO) learners dataset

The LEO dataset is a collection of tables that have been created for the purpose of secondary analysis from employment, benefits, self-assessment and earnings data from the DWP and Her Majesty's Revenue and Customs (HMRC) and DfE data. The dataset is designed to allow users to follow particular Key Stage 4 (KS4) cohorts from the time they finish KS4. DfE are currently preparing for a standardised LEO extract to be made available as a pilot for researcher access through the ONS Secure Research Service (SRS).

Interest in embedding CLS data into LEO

Whilst DfE's priority remains to roll a core service out, senior members of the LEO team in DfE with whom we have spoken agreed that it should also be a medium term strategic priority to enable LEO data to be linked to longitudinal cohort study data and it is their intention to factor this into their strategic programme. In the short-term it may also be of strategic benefit to factor this linkage into the DfE's pilot project with ONS and DfE suggested it could be of value to ONS to use the cohort data as a test case for how onward linking would work. DfE has suggested that CLS contact the ONS SRS to garner interest in this.

Administrative Data Research UK (ADRUK) Strategic Hub

ADRUK are working on a number of themes under its strategic impact programme. Under the *Data for Children* theme ONS is linking educational records to the 2011 Census for those children who were 14 in 2011. This will also be supplemented with the English School Census.

Interest in embedding CLS data into Data for Children and other data

The *Data for Children* data won't match up to CLS cohort data as we don't have CMs who were 14 in 2011. However, if ONS would be willing to do something similar with data that would match up to ours this could be of interest. We don't have consent from CMs to link to Census data so we would need to consider if we could do this. CLS already link National Pupil Database data into their cohort studies and if it's not possible to link to the Census then there is some question around the utility of doing this. However, the exercise could be one around increasing access to, and knowledge of, the cohort data. ADRUK are also have a wider remit to try to bring together different administrative data sources across the UK and to facilitate the linkage of cohort studies into them and so we should continue to liaise with ADRUK over its planned linkages.

ONS Longitudinal Study (LS)

ONS link Census data every ten years to administrative data. This is a 1% sample based on people born on one of four birth dates, the details of which are kept highly secure.

Interest in embedding CLS data into ONS Longitudinal Study

Given the ONS LS sample limitations it is unlikely to overlap with CLS cohort study data and ONS are not able to reveal the sample birth dates for reasons of data confidentiality. It's possible that in future there could be an appetite for a longitudinal population-wide administrative dataset for which there would be greater potential for embedding CLS data but this would need to be driven at a high level within ONS.

Other ONS linked datasets

There is a huge amount of activity within ONS around administrative data linkages, primarily driven by the work being done to see if administrative data could replace the 2031 Census. There's therefore a focus on creating Census-like data out of administrative datasets with data including GP registers, education data, Customer Information System (some limited data from HMRC and DWP), births and deaths and migration data.

Furthermore, the Digital Economy Act is enabling ONS to require other government departments to give them data. They can also compel the private sector too but this is as yet untested. ONS are in the process of receiving more administrative data such as those from UCAS, HESA, DWP and HMRC.

Interest in embedding CLS data into ONS linked datasets

Regarding Census-like linked administrative datasets, conceptually this would give a strong framework to link other data in, such as CLS cohort data. There is a lot of potential for CLS to explore embedding its survey data into the other data that ONS are in the process of acquiring, including UCAS and HESA data (who CLS are also liaising with around linking these records into their data). A way forward would be for CLS to put forward some specific proposals to ONS to go through their Microdata Release Panel and to provide strong use cases that align with policy priorities.

SAIL Databank

The Secure Anonymised Information Linkage Databank includes a large number of administrative data for Wales, particularly around health and education records which can be linked to survey data. The SAIL Databank in Wales uses a population spine for its records linkages and this is based on the NHS GP Registration Scheme so that all the datasets in its collection share an Anonymised Linkage Field (ALF). This greatly facilitates records linkages.

Interest in embedding CLS data into the SAIL Databank

Via the SAIL Databank, CLS have already linked some of the health records into its MCS data and are in the process of updating these linkages. Because SAIL use a population spine with a shared ALF this means that in a sense the data are, or will be, embedded into the SAIL Databank and if our cohort data are listed as one of the

'restricted core datasets'[1] on the SAIL web site this will attract more applications for the data to potentially be used in this way.

Potential challenges

Legal

Legally there should be no barriers for cohort data linkages into linked national administrative datasets as there are a number of applicable legal gateways which allow for the processing of personal data, including those provided through the General Data Protection Regulation, the Statistics and Registration Service act and the Digital Economy Act. Although ONS hold health and social care data for its own purposes it cannot currently make those data available for research purposes under the Digital Economy Act but are looking into how to make the data available in future.

Following guidance from Research Councils, University College London and the Information Commissioner's Officer, at CLS we have determined that our legal basis for processing (acquiring, linking and sharing) personal data is for a public task under the General Data Protection Regulation (GDPR).

Ethics and consent

In addition, for ethical purposes, we currently continue to seek explicit informed consent from CMs to access and link specific administrative records to their survey data on an opt-in basis– these include health, education, economic and crime records.

The consent forms do not explicitly state that CMs' data would be made available linked to multiple data sources at the same time but it is our understanding that they do not preclude us from doing this and we believe that it would be reasonable to expect that we would do this. However, we should aim to review our privacy notices to ensure that this is clear.

The consent forms allow us to share CMs' personal data (names, addresses, dates of birth etc) with the relevant government departments for the purposes of matching provided the personal data isn't supplied alongside the CMs' survey data. The forms also say that we will make the data available through the UK Data Service under appropriate conditions of access. It's our understanding that this should not preclude us from also making the data available through other similar services with equivalent processes around information governance and data security such as the ONS SRS.

However, perceptions around matching into multiple sources of administrative data and their use by government departments would need to be very carefully managed. It is important to avoid any misconceptions and to make it clear that the linked data will not be used by government departments for anything other than statistical and research purposes (i.e. they will not be used to identify, target or make any decisions about any specific individual).

[1] https://saildatabank.com/saildata/sail-datasets/#core-restricted

Overall, for those linkages we have consent for, there shouldn't be any major ethical or consent issues. However, it would be challenging to link to linked population datasets for which we do not have consent to all elements (e.g the example of Census data given above). If we wanted to link into data at the individual level on an unconsented basis this would need to be very carefully managed and a separate review under this project considers some further issues around this and how this could be handled.

Technical

It would be necessary to consider issues around data matching when matching into multiple data sources, such as LEO. Pragmatically we might choose to do a single matching exercise into one of the data sources but we would need to consider any gaps in the data we are matching into and whether we might need a more complex matching exercise, weighing up any additional coverage gains against complexity.

Ideally we would be able to use an administrative data population spine to enable all of our data linkages. The ESRC's Longitudinal Studies Strategic Review in 2017[2] recommended the creation of such a spine from which to sample a new cohort and refresh existing samples and such a spine would also enable easier records linkage and the embedding of cohort studies into administrative data sources. A project led by Bristol University is currently gathering evidence to support ESRC and HDRUK in assessing the feasibility and potential value of a UK population spine (which CLS will feed into) but such a spine is unlikely to be possible in the near future. As mentioned earlier, the SAIL Databank in Wales uses a population spine for its records linkages based on the NHS GP Registration Scheme and this greatly facilitates records linkage particularly since most of the data in SAIL relate to health data. Ideally any UK population spine would use not only health but also economic data sources for the sake of completeness (given that some people will be missing from certain registers/administrative sources).

Methodological

There are issues around consent bias and how to handle unlinked records from an analytical point of view. It would probably not be possible for us to flag any unlinked cases in the embedded dataset where no consent was given to match so, within the embedded dataset, researchers wouldn't have any information about CMs who didn't consent. A body of work would therefore be needed to explore consent bias. This could be done using the full survey data to compare people who consented with those who didn't. Within the embedded data, it would also be possible to look at differences in the administrative data between the linked and unlinked records to give measures of combined survey bias and consent bias and make decisions and recommendations on how this should be handled. This would need to be well documented for researchers.

[2] https://esrc.ukri.org/files/news-events-and-publications/publications/executive-summary-longitudinal-strategic-review-2017/

Establishing scientific value

It's important to establish scientific use cases in advance and to ensure that any issues such as those above have been thought through sufficiently. We would need to demonstrate the scientific value and benefit of linking the data in this way so that the results show a public benefit and can positively inform future commissions for other data linkages. If the data aren't used in this way then this could impact on the success of requests for future data linkages.

Onward data sharing

In terms of the mechanism for researchers to access the linked data there are two suggested options which would need to be further discussed and agreed:

- For data where the administrative data are linked into the cohort data it would make sense to continue to deposit these standalone datasets with the UKDS. The UKDS is the established route for CLS's survey data and would avoid fragmentation of access across different services. The UKDS precedent has been set for NPD data linked into the cohort data but has not yet been set for HMRC and DWP data, although good progress is being made with the former and CLS are in the early stages of negotiation with the latter.
- For data where the cohort data are linked/embedded into the administrative data it would make sense for these data to be made available through the ONS SRS as this is where the vast majority of administrative data will eventually be held and accessed and adding cohort data would provide richer detail for a subset of records. It is also highly unlikely that full administrative data would ever be deposited with UKDS.

It was suggested that it might be prudent to consider a staged approach to embedding survey data so that they are initially made available just linked into the cohort studies for a period of time in order to establish their value and use and feed into the case to data controllers to make available the wider dataset into which the cohort data are embedded. However, CLS has already linked health and education data into its cohort data and can demonstrate the value and use of data linked in this way and is therefore already in a good position to make strong scientific use cases for the wider datasets.

Recommendations

There is strategic interest in embedding cohort data into linked administrative datasets and CLS should continue working with ONS, ADRUK and relevant government departments and organisations across the UK (the coverage of the different CLS cohort studies ranges from England only, to GB to UK) to enable this to happen. In particular CLS should, subject to capacity and funding constraints -

- Make further contact with ONS about being involved in the LEO pilot dataset to be made available through ONS SRS.
- Discuss with the ADRUK board the possibility of widening the *Data for Children* to include a different age group so that the data can be linked to

MCS data, and to keep abreast of wider strategic developments within ADRUK (and noting CLS Director Alissa Goodman is well placed to do this as a member of the ADRUK research commissioning board (RCB) a member of the Data for Children Themed partnership, and its Expert Group on ADRP and support for Longitudinal and Cohort Studies).

- Explore further whether it's possible to link cohort data to Census data.
- Produce scientific use cases for embedding cohort data into linked administrative datasets to convince data controllers of the utility of doing this.
- Explore the differences between people who consent and those who don't in its cohort studies with an aim to making recommendations about how any potential consent bias could be handled for embedded data.
- Continue to work on depositing standalone datasets with administrative data linked into cohort data but with a view to making wider administrative data with cohort data embedded available.
- Continue to work with SAIL in Wales on refreshing data linkages and also explore the option of embedding cohort data with Scottish and Northern Irish government departments (where coverage allows).
- Review its privacy notices for cohort members and information provided to reflect how the data may be used and where they may be accessed.
- Consider practical ways to work with ONS to enact data linkages, including CLS staff secondments.
- Keep abreast of and feed into the scoping study for a UK population spine.

Annex – who we spoke to

Thanks to the following people for taking the time to speak to us. Any errors in interpretation are the author's own.

ESRC

- Andy Boyd – ALSPAC Data Linkage & Information Security Manager & Lead for ESRC-HDRUK Admin Spine Scoping Study

ONS

- Iain Bell - Deputy National Statistician and Director General for Population and Public Policy
- Rose Elliot - Head of Strategic Data Curation, Admin Data Research
- Ben Humberstone - Deputy Director for Health Analysis and Life Events
- Jim Newman - Head of Longitudinal Study Development, Centre for Ageing and Demography, Public Policy Analysis
- Jason Riches – Legal and International Services
- Karen Tingay - A Principal Statistical Methodologist (formerly at SAIL)

SAIL

- David Ford - Professor of Health Informatics, Co-Director of SAIL
- Ronan Lyons - Professor of Public Health, Co-Director of SAIL

DfE

- Clare Baker – Head of Performance Tables Development Unit, Education Data Division
- David Burnett – Lead for delivery of the LEO programme
- Natalie Masters – Education Data Division

DWP

- Mike Daly – Central Analysis Division, oversight of DWP's research programme