EDUCATION EVIDENCE BASE

Submission to the Productivity Commission National Education Evidence Base Inquiry

May 2016

The Population Health Research Network (PHRN) welcomes the opportunity to comment on the National Education Evidence Base: Issues Paper.

The PHRN's submission provides a brief introduction about the PHRN and responds to some questions from the issues paper.

About the Population Health Research Network

In Australia, information about an individual's health, education and welfare is recorded throughout their lives as they come in contact with service delivery organisations and agencies, including hospitals (public and private), health departments, schools and other government agencies. The collection of this data is often required under legislation and the information is stored in secure computer databases within the responsible agencies.

Data linkage is a technique for creating links between pieces of information that are thought to relate to the same person, family, place or event (Hobbs & McCall 1970). This function is often performed by specialist data linkage units which facilitate access to linked data to enable research for the public benefit.

Australia has been at the forefront of the development of methods to provide researchers access to linked data whilst preserving privacy since the establishment of the Western Australian Data Linkage System (WADLS) in 1995 and the Centre for Health Record Linkage (CHeReL) in New South Wales (NSW)/Australian Capital Territory (ACT) in 2006 (Holman et al 2008; Lawrence, Dinh & Taylor 2008).

The Population Health Research Network (PHRN) commenced in 2009 and is funded by the National Collaborative Research Infrastructure Strategy. The University of Western Australia is lead agent for PHRN. The PHRN's primary purpose is to build and support the operation of collaborative, nationwide data linkage infrastructure capable of securely and safely linking data collections from a wide range of sources including within and between jurisdictions and across sectors. Australia now has the facilities and capabilities to link and provide access to linked data in all jurisdictions. PHRN achievements include:

- Establishment of new data linkage units in Queensland, Victoria, Tasmania and South Australia;
- Establishment of accredited Commonwealth Integrating Authority at the Australian Institute of Health and Welfare;
- Establishment of a unit to undertake cross-jurisdictional linkage at Curtin University;
- New online application and secure data delivery systems which facilitate access to data; and
- Remote access data laboratory (SURE) that enables researchers to access linked datasets in a secure environment from anywhere in Australia.

For more information, please visit http://www.phrn.org.au/

Who can use Australia's existing national data linkage infrastructure?

Australia already has a national data linkage system which should be leveraged for delivery of high quality, linked education data. It is not restricted to health data. A number of education data collections including the Australian Early Development Census, NAPLAN and school enrolment data are already included in state/territory data linkage systems. If linkage variables are available there is no technical barrier to the linkage of other education data collections.

What are the main challenges and impediments to implementing data linkage in the education sector?

The legal framework for access to publically funded data for research

Australia has a complex legal framework governing the collection, use and disclosure of data for research. There is also variation between jurisdictions and in the clarity and terms of individual legislation. Whilst in most cases the empowering legislation permits the use and disclosure of data for research, provisions in the various education statutes may limit access to education data and each research project may need to be considered on a case by case basis.

Data custodians' primary responsibility is to ensure that they comply with the law when considering requests for access to data for research. They may not always feel confident about decision making in such a complicated legal environment and may act cautiously.

There are a number of approaches that could overcome the barriers that the legal framework causes. In the short term the provision of guidance materials and training to assist data custodian agencies to process requests for access to data would be worthwhile. In the medium term changes to legislation to clarify the use of specific data collections for research may be required. In particular, each jurisdiction (state, territory and Commonwealth) should have legislation that covers the collection, use and disclosure of education information. In the longer term Australia should consider a more uniform national approach.

Privacy legislation

Australia's data protection or privacy legislation protects the right to information privacy by limiting the use and disclosure of personal information without consent. Australia's privacy legislation does provide for the disclosure and use of personal information for public health research. However, it does not cover disclosure of personal information for education research (Adams & Allen 2014). Amendments to Commonwealth, state and territory privacy legislation to cover use of personal information for education research could be considered.

Commonwealth data

Australia has education data collections and the national infrastructure to safely and securely link Commonwealth and Commonwealth/state/territory data to provide information that will inform development of education policy, monitor policy implementation and measure educational outcomes. Despite this capacity, significant barriers to access Commonwealth linked data remain. The selection of education and training data collections listed in Table 1 summarises the range of national data collections available in Australia but apart from a few collections, there is limited information in the public domain on how to obtain approval to access data as well as the limited extent of linked Commonwealth education data resources. This limits the ability of government and researchers to take a population level approach to linked data research to measure educational outcomes.

With respect to linked Commonwealth data resources, it can take many months to link data from the large Commonwealth data collections for a specific project and these links are generally destroyed when the project is complete. All states and territories now have enduring linkage keys with at least 10 years of linked health data. There are not enduring links between Commonwealth data collections. Enduring linkage between Commonwealth, state and territory health data collections is also rare. Allen et al (2013) suggest that there is a risk-averse culture in Commonwealth departments which focuses on privacy risks and may not place sufficient weight on the benefits of the research findings and the risks of not doing the research.

Changing data management practices is not trivial. Continuing support for a national process to prioritise and address the variations over the short to medium term is suggested.

State/Territory data

Education and training data collections in states and territories are diverse and generally not well documented. All PHRN state/territory data linkage units have AEDC data and a number have education and other human services data.

Each jurisdiction should have documented administrative arrangements for considering requests for access to the information (if this is not covered in legislation and any related regulations). As far as possible, access arrangements should be standardised within a jurisdiction and across jurisdictions.

It may be cost effective to manage education data from a number of agencies as a warehoused resource within the Commonwealth and within each state/territory.

Priorities

- Data custodians to prepare and publish metadata on key data collections (1-2 years)
- Standardisation of processes to consider requests for data access within and between jurisdictions (1-2 years)
- Standardising data collection practices should continue (ongoing)
- Expansion of education data collections included in jurisdictional master linkage keys (ongoing)
- Jurisdictional data warehouses for administrative content data to be progressed (2-3 years)
- Legislation review should commence at an early stage as it will take time to achieve change (1-3 years).

What are the relative advantages and disadvantages of using probabilistic or deterministic linkage techniques to link datasets?

Probabilistic and deterministic linkage are two approaches to data linkage, which are documented in the issues paper (Box 6). Deterministic linkage involves exact matching of linkage variables such as name, address, date of birth and sex, a subset of these variables or a unique identifier. That is a pair of records would be considered a match if the agreed identifiers are identical. Probabilistic linkage uses a combination of linkage variables such as name, address, date of birth and sex which are weighted to determine links. Deterministic linkage requires assumed knowledge about the quality of all linking variables whereas probabilistic linkage is able to adjust for name and address changes because this information is generated and tolerated (Herzog, Scheuren, Winkler 2007). The implementation of probabilistic linkage takes a significant more amount of time compared to deterministic linkage (1 min vs 2 mins to 2 hours for a simulated scenario) (Zhu et al 2015).

PHRN data linkage units may offer both deterministic and probabilistic methods of linking records. Simulated studies have reported that probabilistic linkage is superior to deterministic linkage in all scenarios because of its accuracy with data of varying qualities and its transparency and flexibility (Zhu et al 2015; Tromp et al 2011). Deterministic linkage missed more matches which detracts from the quality of the linkage (Tromp et al 2011).

What are the costs and benefits of expanding the Unique Student Identifier national to students in schools and early childhood education and care?

Reliance on the Unique Student Identifier, or any unique identifier, may limit intra and cross-jurisdictional research because:

- the use of a Unique Student Identifier will not solve the problem of linking education data to data collections outside of the education sector such as births, deaths, hospital, justice and housing
- there are legislative and ethical barriers to using an identifier created for one purpose which is then used for another purpose.

In general, probabilistic linkage using identifying variables such as name, address, date of birth and sex provides better linkage quality that a Unique Student Identifier when linking across years, geographical locations and data collections.

What lessons can be learned from data access arrangements in non-education sectors and in other countries?

Documented and publicly available information

Data access arrangements which are well documented and available in the public domain is key. The PHRN facilities and services have documented data access arrangements available in the public domain (e.g. PHRN website). These facilities and services could be leveraged and documented information about core education data collections could also be available on the PHRN website.

Overseas data linkage units in the UK (SAILDatabank and the Farr Institute) and Canada (Population BC and Manitoba Centre for Health Policy) all have documented data access arrangements. For more information see:

SAILDatabank http://www.saildatabank.com/

PopDataBC https://www.popdata.bc.ca/

Manitoba Centre for Health Policy http://umanitoba.ca/faculties/health_sciences/medicine/units/community_health_sciences/departmental_units/mchp/

Farr Institute http://www.farrinstitute.org/

PHRN facilities and services

The PHRN provides a number of national eResearch tools and services to assist researchers to access linked data efficiently and securely.

The PHRN Online Application System has been developed to improve the efficiency of the application process for single state and cross-jurisdictional linked data projects. The PHRN Online Application System can be used to apply for linked education data. The unified form reduces the number of application forms required and enables researchers to submit their applications simultaneously and track their applications online. The PHRN Online Application System has been endorsed by all PHRN data linkage units as the agreed application process for cross-jurisdictional linked data projects. For more information visit http://www.phrn.org.au/for-researchers/data-access/online-application-system/

The Secure File Exchange Service (SUFEX) is one of the options that data custodians can use to send files to the PHRN data linkage units and researchers. SUFEX uses a

secure online application that allows users to send and receive files from anywhere at any time. For more information visit http://www.phrn.org.au/for-researchers/services-for-researchers/sufex/

The Secure Unified Research Environment (SURE) is Australia's first and only remote-access data research laboratory purpose-built for the analysis of linked, routinely collected data. Researchers must use the SURE to access and analyse linked Commonwealth and state/territory data files for approved research projects in Australia. For more information visit http://www.phrn.org.au/for-researchers/services-for-researchers/secure-unified-research-environment-sure/

Information Agreements

Australia has had a National Health Information Agreement since the early 1990s, also a National Health Data Dictionary supported by national committee processes and an online metadata repository at the AIHW. The education sector may wish to consider if a similar national system for education data may be achievable.

In the event of conflict between data users and data managers are there effective dispute resolution mechanisms?

There are currently no dispute resolution mechanisms between data users and data managers. If a researcher has been refused access to government data or their access has been excessively delayed, they are unable to seek independent external review of a data custodian's decision. It has been suggested that lodging a complaint with the Commonwealth Ombudsman on the basis that a data custodian acted "wrongly, unjustly, unlawfully or unfairly" may be one dispute resolution pathway but this would be difficult to establish (Adams & Allen 2014). Furthermore, there is no dispute resolution mechanism when failure to release data for research has led to harm (Allen et al 2013).

References

Adams C, Allen J. 2014. Government databases and public health research: Facilitating access in the public interest. *Journal of Law and Medicine*. 21(4): 957-972.

Allen J, Holman CD, Meslin EM, Stanley F. 2013. Privacy protectionism and health information: Is there any redress for harms to health? *Journal of Law and Medicine*. 21(2): 473-85.

Herzog TN, Scheuren FJ, Winkler WE. 2007. *Data quality and record linkage techniques. Springer*, New York.

Hobbs MS, McCall MG. 1970. Health statistics and record linkage in Australia. *Journal of Chronic Diseases*. 23(5): 375-81

Holman CD, Bass AJ, Rosman DL, Smith MB, Semmens JB, Glasson EJ, Brook EL, Trutwein B, Rouse IL, Watson CR, de Klerk NH, Stanley FJ. 2008. A decade of data linkage in Western Australia: Strategic design, applications and benefits of the WA data linkage system. *Australian Health Review*. 32(4): 766-77.

Lawrence G, Dinh I, Taylor L. 2008. The Centre for Health Record Linkage: A New Resource for Health Services Research and Evaluation. *Health Information Management Journal*. 37(2): 60–62.

Tromp M, Ravelli AC, Bonsel GJ, Hasman A, Reitsma JB. 2011. Results from simulated data sets: Probabilistic record linkage outperforms deterministic record linkage. *Journal of Clinical Epidemiology*. 64: 565-572.

Zhu Y, Matsuyama Y, Ohashi Y, Setoguchi S. 2015. When to conduct probabilistic linkage vs deterministic linkage? A simulation study. *Journal of Biomedical Informatics*. 56: 80-86.