Curtin University

Productivity Commission

Data Availability and Use

Submission from

Centre for Data Linkage
Curtin University

July 2016

This page has been left blank intentionally.

Submission to the Productivity Commission

Data Availability and Use

Collection, storage and management of data have come a long way in the last two decades with significant developments in technology in terms of power and capacity. The volume of digital data is increasing exponentially, providing more available data for research. This increase in information also includes routinely collected administrative data which provides the building blocks for critical analysis used to shape policy, evaluate performance and improving public services.

Private and public organisations collect significant amounts of data on their business processes and services. Analysing these individual and aggregate records provide an opportunity to improve knowledge of the environment. Data manipulation and analytics can unlock the potential within the data and combining data from different sources can often provide a better understanding of the 'big picture'.

"Data linkage" is a way of bringing data together to provide information on the whole population, generating a more complete picture of the community than is possible using other research methods. It is also a very cost-effective research tool.

The Centre for Data Linkage (CDL) is a national leader in the research and development of technology to safely and securely maximise the value of data available for research. The CDL was established in 2009 within Curtin under the National Collaborative Research Infrastructure Strategy (NCRIS) and is a member of the Population Health Research Network (PHRN). The focus was to develop and implement secure, state-of-the-art national infrastructure to enable cross-jurisdictional data linkage for research. As part of the project, the CDL has undertaken research into both technical and methodological aspects of data linkage to improve performance and capacity. These have been incorporated into the design and construction of efficient linkage infrastructure that maintains data integrity and security. The CDL has also provided technical advice and support to linkage units across Australia (PHRN). This support and guidance includes organising and running technical forums to discuss and share common challenges in designing, building and operating data linkage infrastructure. Since 2009, the CDL has:

- Designed, built and operated a secure environment to host the CDL data linkage infrastructure (including scalable Cross Jurisdictional linkage capabilities across Australia (1-4));
- Undertaken research into linkage systems, methods and models (including evaluation of linkage products and systems; development of Quality Assurance tools; a review of, and research into, data cleaning and standardisation; a review of privacy preserving data linkage techniques and a number of other on-going research projects). The quality of that research is evidenced through numerous publications in peer-reviewed journals (see Attachment A for list of publications);
- Adopted innovative approaches to data linkage functionality, performance, algorithms (including scalability in matching algorithms), parallel processing and database optimisation. The CDL has developed data linkage infrastructure that is reflective of

contemporary standards in IT and incorporates latest technologies in data linkage. These include, for example, a large, scalable linkage system with concurrent processing; multi-threading; differing grouping strategies; the ability to undertake project and enduring linkages; Privacy Preserving Record Linkage (PPRL) options; and

- Provided technical advice and assistance – includes direct assistance, customised training of technical staff, information sharing sessions with other PHRN members, publication and broader distribution of reports, promotion and hosting of PHRN Technical Forums. Technical leadership and innovation are evidenced through a variety of published articles on data linkage methods (5-10).

Questions on high value public sector data

What public sector datasets should be considered high-value data to the: business sector; research sector; academics; or the broader community?

University based researchers in Australia are recognised as world leaders in the use of public sector data for research. Health, education and criminal justice datasets have been widely used to gain a better understanding of systems and services. Using a whole system approach recognises that many interacting factors can influence individual parts of the system and that solutions to problems have to be developed taking these factors and interactions into account. The demand for data from these sectors is strong and growing.

Research using routine administrative data in Australia and overseas has demonstrated its value, access to health, education and criminal justice datasets are crucial in supporting policy and improving public services.

What characteristics define high-value datasets?

Many of the administrative datasets are collected by government departments and other organisations to measure and monitor operations during the delivery of a service. The high-value characteristics of administrative data include coverage of the whole population, which allows analysis of small group and vulnerable, collection quality standards (with metadata) and long-term analysis to a level of detail not permitted by sample surveys. The use of administrative data provides a cost effective and efficient method for population based research and avoids imposing a further burden on respondents.

What benefits would the community derive from increasing the availability and use of public sector data?

The benefits of data sharing have been shown to improve research skills and analytical tools significantly for complex integrated data, enabling new research that enhances the delivery of public services. Access to high quality information is essential for efficient and effective government systems and services.

Questions on collection and release of public sector data

What are the main factors currently stopping government agencies from making their data available?

Government and University departments in Australia understand that administrative data can provide an unparalleled resource for the monitoring and evaluation of services.

However, for a number of reasons, these data have not been accessible to researchers. An additional barrier in Australia is that health data are collected by different levels of government – thus not all available through any one authority.

Barriers to data access

The main challenges or barriers to be resolved in realising the potential health and health related data in Australia include:

- The distributed nature of health care responsibilities, coupled with a federated legal system, means that any long-term solution requires cooperation between State and Commonwealth stakeholders;
- Authorising environments - it is time-consuming to establish projects in terms of approvals and governance arrangements. Establishing transparent and consistent procedures that manage/streamline all the processes involved in data access would ensure effective and efficient use of information. Transparent and consistent processes would reduce uncertainty around ethics, privacy and data custodian constraints;
- Legislation – many of the significant Commonwealth and State datasets are subject to specific legislation that defines the conditions of data release and/or use. The extent of this type of legislation and its complexity creates difficulties of interpretation with regard to the release of data for research projects;
- Operational efficiencies – The "quantity of data" emerging from electronic health collections also poses challenges (i.e. Big Data). Increasing demand on data linkage services also puts significant pressure on infrastructure to deliver in a timely fashion;
- Capacity – at present, the operations required to fulfil a data request can pose a substantial burden on organisational resources. Infrastructure enabling data linkage needs to be scalable, fast and efficient to ensure timely responses to important policy and research questions;
- Expertise – data linkage requires expertise in three broad areas: knowledge of the datasets available for linkage along with their characteristics and limitations, skills in linkage methods and skills in using/analysing linked information. By itself, a basic-level ability to use available linkage software is insufficient, because correct interpretation of linked datasets depends on an understanding of the structure and content of, and variation within the component datasets;
- The funding environment - The PHRN represents a major co-investment by the Commonwealth Government and PHRN partners in national data linkage infrastructure. However, the current funding model is inadequate (time-frames too short; uncertainties high) which makes operation, maintenance and support of the infrastructure difficult and innovation virtually impossible. Without long term funding the infrastructure will be unable to realise its full potential.

How could governments use their own data collections more efficiently and effectively?

Efficiencies could be gained through:

- Cooperation: Development and endorsement of agreed principles/statements asserting value of data for public benefit and supporting the release of data for research;

- Enhance data flows: Exploring and implementing effective methods of enabling data flow (especially for complex, multi-dataset, multi-agency national or cross-jurisdictional projects). Including agreed data flows which provide comprehensive security and make collaborations between researchers easier and more efficient;
- The development of the eHealth Record Systems (My Health Record) through the National E-Health Transition Authority (NEHTA) also provides opportunities for secondary use of health data for government and university research through data sharing and linkage with other information sources;
- Interoperability: Australia needs to ensure that infrastructure and technologies are interoperable and responsive to environmental changes around legislation, information technology, security and privacy. Common platforms allow the transfer of expertise, learning and skills between government and university teams.

Should the collection, sharing and release of public sector data be standardised? What would be the benefits and costs of standardising? What would standards that are 'fit for purpose' look like?

- Streamlined access: Creating a streamlined and consistent application and approval processes for research projects (especially for complex national/cross-jurisdictional projects using health and health related data). At present, approvals processes are numerous and lengthy. Developing a national, co-ordinated approach to ethics applications and approvals is required to expedite access (simplify the process; reciprocal/mutual recognition);
- Transparency and public accountability: Clear processes around assessments (for example, the balance of public good against the privacy imposition and risks to confidentiality) are essential as is public accountability. Accountability mechanisms could include the creation of an independent auditing or oversight body, community representation on steering committees or an advisory committee.

What specific government initiatives (whether Australian Government, state, territory or local government, or overseas jurisdictions) have been particularly effective in improving data access and use?

Unfortunately, it seems that all government agencies (national and international) seem unable to share unit data for research. Progressive policies, with suitable safeguards, around data sharing for research are required to maximise the value of collected data.

Questions on data linkage

The methods and techniques around data linkage in Australia are well established, and the new developments (exploiting advances in technology) have the potential to improve timeliness and efficiency. Leveraging these developments will fast track the research and policy making programme.

Which datasets, if linked or coordinated across public sector agencies, would be of high value to the community, and how would they be used?

Health, education and criminal justice datasets provide a stable platform for both government and university research teams to gain a better understanding of population interactions with systems and services.

Which rules, regulations or policies create unnecessary or excessive barriers to linking datasets?

Given the federated nature of health care service delivery in Australia (i.e. some services are delivered and administered at State level, while others are delivered and administered at national or "Commonwealth" level), the impact of Commonwealth funding, serving planning and health outcomes can only be achieved through efficient cross-jurisdictional and national infrastructure. Experiences from other countries demonstrate the need to harness and harmonise the power and experience of linkage services and systems to improve the efficiency and quality within overall data linkage infrastructure.

- Authorising environments - it is time-consuming to establish projects in terms of approvals and governance arrangements. Establishing transparent and consistent procedures that manage/streamline all the processes involved in executing a linkage project (end to end arrangements) would ensure effective and efficient data linkages. Transparent and consistent processes would reduce uncertainty around ethics, privacy and data custodian constraints;
- Legislation – many of the significant state and commonwealth datasets are subject to specific legislation that defines the conditions of data release and/or use. The extent of this type of legislation and its complexity creates difficulties of interpretation with regard to the release of data for linkage projects;

How can Australia's government agencies improve their sharing and linking of public sector data? What lessons or examples from overseas should be considered?

Data linkage in the United Kingdom is undergoing a significant expansion of capabilities. Charities, Research Councils (the Medical Research Council and the Economic and Social Research Council), Government and other bodies have invested over £200million in the new Farr Institute - a collaborative 'partnership model' between government and the university sectors. The aim of the Farr Institute is to provide an integrated research platform for health and other Government sectors. Major centres are located in London, Dundee, Manchester and Swansea and link research in 19 universities across the UK and Northern Ireland.

The Farr Institute supports safe use of patient and research data for medical research across all diseases in the UK. Its research supports innovation in the public sector and industry leading to advances in preventative medicine, improvements in healthcare and better development of commercial drugs and diagnostics. The Farr Institute will also provide new insights into the understanding of causes of ill health which in turn will guide new biomedical research discovery. In preparation for these national developments, data linkage experts from Australia have provided advice and support to various Farr Institute nodes.

Legal, administrative and technical issues across the world have impacted on the ability to undertake linkage of particular datasets. New record linkage techniques, collectively referred to privacy-preserving record linkage, significantly reduce privacy risks as they operate on de-identified information and do not require the release of personal identifiers. Researchers from Australia, Germany, Canada and the United Kingdom are developing software that implements Privacy Preserving Record Linkage (PPRL) for use in operational record linkage settings. Adoption and application of these methods would increase capabilities and enable linked research opportunities as additional datasets are made available through a PPRL framework.

With significant international investment in data linkage and ‘Big Data’ science (supporting a push for open government) in the United Kingdom and Canada, long term funding of data linkage infrastructure in Australia is required to avoid losing the competitive advantage that Australia has gained in the international data linkage arena and in the fields of research that use linked data.

Questions on resource costs of access

How should the costs associated with making more public sector data widely available be funded?

Improved funding model. The funding environment is necessary to enable improvements to and expansion of services and delivery to a variety of user groups; to assist in prioritisation of activities. Without long term funding the infrastructure will be unable to realise its full potential. (Short-term planning/funding makes operation, maintenance and support of the infrastructure difficult and innovation virtually impossible).

Is availability of skilled labour an issue in areas such as data science or other data-specific occupations? Is there a role for government in improving the skills base in this area?

Expertise – data analytics requires expertise in three broad areas: knowledge of the datasets available for research (along with their characteristics and limitations), skills in data manipulation methods and skills in using/analysing information. By itself, a basic-level ability to use available data is insufficient, because correct interpretation of datasets depends on an understanding of the structure and content of, and variation within the component collections.

Programs like the NSW Biostatistician Training Program, established in 2000, provides broad training that enables graduates to apply biostatistical expertise to many different domains of public health practice. Graduates are skilled up to work as biostatisticians in a range of public health services, research, development, policy and planning positions.

Questions on privacy protection

What types of data and data applications (public sector and private sector) pose the greatest concerns for privacy protection?

All university based research projects require ethical, custodian and institutional approval before they can proceed. At each stage, privacy and confidentiality restrictions add to the project-specific governance framework and control arrangements. The scope of the Information Governance Framework is to provide a systematic approach to safeguarding all sensitive information involved in the research activities. As a result, secure research facilities provide a safe environment to perform analysis on de-identified and/or appropriately confidentialised datasets.

How can individuals’ and businesses’ confidence and trust in the way data is used be maintained and enhanced?

Data custodians, researchers and record linkage centres have worked together to develop data access and usage models that comply with information privacy laws and provide necessary guards to privacy e.g. Australian Government High Level Principles for Data Integration (11). Moreover, record linkage units have implemented an array of best practice data governance policies to minimise the risk to privacy posed by their operations (1, 12-16).

Project-specific information governance encompasses people, processes, information technology (IT) systems, information and physical assets that support the research activities.

What weight should be given to privacy protection relative to the benefits of greater data availability and use, particularly given the rate of change in the capabilities of technology?

In an era of 'big data' development, there are significant challenges around data sharing and linkage. These include caution around data sharing and linkage and conservative interpretation of legislation around data release. There needs to be more thought given to the balance between data access, privacy and public benefit in research.

Are further changes to the privacy-related policy framework needed? What are these specific changes and how would they improve outcomes? Have such approaches been tried in other jurisdictions?

In most Western countries, information about an individual's health and welfare is collected as they come into contact with service delivery organisations and other government agencies, including hospitals (public and private), health departments and other human or social service authorities (e.g. education, criminal justice). Over time this data accumulates, providing a rich store of information that can be used to inform policy making and improve the health and social status of the entire community.

Technological advances have improved the accessibility, quality and integration potential of this data for research. In parallel, these 'big data' developments have helped establish flexible and transparent governance models that balance both privacy and the public interest in research. Developing proportionate governance frameworks based on clear guiding principles allows accurate assessment of risks associated with data use/sharing/linkage and assigns appropriate safeguards (17).

How could coordination across the different jurisdictions in regard to privacy protection and legislation be improved?

Many of the significant Commonwealth and State datasets are subject to specific legislation that defines the conditions of data release and/or use. The extent of this type of legislation and its complexity creates difficulties of interpretation with regard to the release of data for research projects. A truly consistent and transparent approach to data access and research assessment is required to ensure equity of data access.

How effective are existing approaches to confidentialisation and data security in facilitating data sharing while protecting privacy?

Existing approaches to confidentialisation and data security are often project specific and often restrictive. The governance of data for research needs to be simplified to allow agile responses to research and policy questions.

Questions on data security

Are security measures for public sector data too prescriptive? Do they need to be more flexible to adapt to changing circumstances and technologies?

Security, ethics and privacy – in addition to legal requirements, access to many Commonwealth and State health datasets are subject to privacy and ethical review. The processes necessary to address privacy and confidentiality concerns are not always transparent. It should be clear what governance process, protocols and standards are required to enable safe and secure access to research data. In addition, the requirement for multiple ethics approvals (often in different application forms) adds additional layers of bureaucracy within the project approval process.

Summary

Overall, data linkage infrastructure in Australia is recognised internationally for its high level of accuracy (linkage quality standards) and innovative technologies/methodologies. The challenge is to realise the potential of the infrastructure currently available across government and university sectors through compatible, sustainable and effective models which can maximise the capacity across all these systems.

Experiences from other countries demonstrate the need to harness and harmonise the power and experience of linkage services and systems to improve the efficiency and quality within overall linkage infrastructure.

Attachment A - References

1. Boyd JH, Ferrante AM, O'Keefe CM, Bass AJ, Randall SM, Semmens JB. Data linkage infrastructure for cross-jurisdictional health-related research in Australia. BMC health services research. 2012;12(1):480.
2. Boyd JH, Randall SM, Ferrante AM, Bauer JK, McInneny K, Brown AP, et al. Accuracy and completeness of patient pathways–the benefits of national data linkage in Australia. BMC Health Services Research. 2015;15(1):312.
3. Spilsbury K, Rosman D, Alan J, Boyd J, Ferrante A, Semmens J. Cross border hospital use: analysis using data linkage across four Australian states. The Medical journal of Australia. 2015;202(11):582-6.
4. Diana Rosman, Katrina Spilsbury, Janine Alan, Anna Ferrante, Angela Young , Emma Fuller, et al. Multi-jurisdictional linkage in Australia: Proving a concept. Australian and New Zealand Journal of Public Health. 2014.
5. Ferrante A, Boyd J. A transparent and transportable methodology for evaluating Data Linkage software. Journal of Biomedical Informatics. 2012;45(1):165-72.
6. Randall SM, Ferrante AM, Boyd JH, Semmens JB. The effect of data cleaning on record linkage quality. BMC Medical Informatics and Decision Making. 2013;13(1):64.
7. Randall SM, Ferrante AM, Boyd JH, Semmens JB. Privacy-preserving record linkage on large real world datasets. Journal of biomedical informatics. 2013.
8. Randall SM, Boyd JH, Ferrante AM, Bauer JK, Semmens JB. Use of graph theory measures to identify errors in record linkage. Computer methods and programs in biomedicine. 2014;115(2):55-63.
9. Randall SM, Brown AP, Ferrante AM, Boyd JH, Semmens JB. Privacy preserving record linkage using homomorphic encryption. Population Informatics for Big Data, Sydney, Australia. 2015.
10. Boyd JH, Guiver T, Randall SM, Ferrante AM, Semmens JB, Anderson P, et al. A Simple Sampling Method for Estimating the Accuracy of Large Scale Record Linkage Projects. Methods Inf Med. 2016;55(3):276-83.
11. Australian Government. High Level Principles for Data Integration involving Commonwealth Data for Statistical and Research Purposes. In: Cross Portfolio Statistical Integration Committee (CPSIC), editor. Canberra: Australian Government; 2010.
12. Lawrence G, Dinh I, Taylor L. The Centre for Health Record Linkage: A New Resource for Health Services Research and Evaluation. Health Information Management Journal. 2008;37(2):60-2.
13. Harris J, editor Next Generation Linkage Management System. Sixth Australiasian Workshop on Health Informations and Knowledge Management; 2013; Adelaide, Australia: Australian Computer Society.
14. Trutwein B, Holman D, Rosman D. Health Data Linkage Conserves Privacy in a Research-Rich Environment. Annals of Epidemiology. 2006;16(4):279-80.
15. Ford D. The SAIL Databank: building a national architecture for e-health research and evaluation. BMC health services research. 2009;9(1):157.
16. Roos LL, Brownell M, Lix L, Roos NP, Walld R, L M. From health research to social research: Privacy, methods, approaches. Social Science and Medicine. 2008;66(1):117-29.
17. Laurie G, Sethi N. Towards principles-based approaches to governance of health-related research using personal data. European journal of risk regulation: EJRR. 2013;4(1):43.