Public sector data

What constitutes high-value public sector datasets – characteristics and potential benefits?

High-value public sector datasets are those that present potential for use and re-use for public benefit. In health, these include datasets that are collected routinely by governments, healthcare providers and third party payers. These datasets are essentially a by-product of operating health services, and those generated to meet regulatory requirements (e.g. births, deaths, health practitioner registration), or specifically to monitor health and to inform disease control and health improvement efforts (e.g. cancer notifications, communicable disease notifications, perinatal data, device registries). These datasets generally have high levels of completeness and accuracy, having rigorously applied standardised data definitions.

Potential benefits of these data for research and policymaking include[1]:

- **Population reach:** many public sector data have whole-of-population coverage, and can be used to study rare outcomes (e.g. adverse events) and population subgroups (e.g. Indigenous Australians). They maximise power to identify even marginal shifts in practice that have resulted from new policies or programs.
- **Longitudinal:** when linked internally or across datasets, these data have a longitudinal structure that supports studies across the life course, enabling long-term follow-up.
- **Avoid nonresponse, attrition and reporting bias:** routine data collections are not subject to the challenges of surveys such as high and rising rates of socially and health-patterned nonresponse and attrition, as well as social desirability, reporting and recall biases.
- **Cost-effective:** the use of routine data for research and evaluation increases return on investment for the public resources expended in collecting them, and studies of whole populations over many decades can be undertaken time-efficiently and cost-efficiently compared with prospective data collection.
- **Real world:** routine data often present the only way to evaluate the outcomes of care in population groups for which there is no evidence from clinical trials, or to evaluate the impacts of policies or services that have been rolled out in a nonrandomised manner.
- **Privacy protection:** proven methods to protect the privacy of individuals and organisations have been successfully implemented in Australia for many years.

Which specific government initiatives have been effective in improving data access and use?

State-based health data linkage units including the Western Australian Data Linkage System (since 1995) and NSW/ACT Centre for Health Record Linkage (since 2006) have increasingly made linked health data available for research.

More recently the Population Health Research Network has contributed to building further infrastructure for linkage of population health data including a data linkage (integration) facility within the Australian Institute of Health and Welfare (AIHW) that is accredited to link Commonwealth data, and a secure remote access facility for analysis of linked data (SURE, see https://www.saxinstitute.org.au/our-work/sure/). However, real improvements in access to and use of linked Commonwealth health data have yet to be realised, with researchers typically waiting for around two years to receive these data for their time-limited grant-funded research.

Which datasets should be linked across public sector agencies and what are the impediments to linking datasets?

The increasing complexity of healthcare in terms of rapidly evolving and fragmented service delivery models, the disruptive impacts of new therapies and technologies, and people living longer with multiple health conditions means that traditional methods guiding health policy and practice are no longer adequate. Linkage of population-based health data provides a platform for answering questions about access to and outcomes of health interventions, health system performance, policy impact and economic analysis. Linkage of health data with 'health-related' public sector data such as data from aged care, early childhood services, community services, justice, housing and the environment allows evidence generation to support cross-sectoral 'joined-up' policymaking.

The foremost impediment is the challenge of bringing together Commonwealth and State-based data collections to generate the most up-to-date evidence about the full-spectrum of health care. For example, it is not possible to quantify the outcomes derived from the Commonwealth's investment in medical and pharmaceutical interventions without linking data across Commonwealth and State boundaries. At a minimum, we require timely linkage of Commonwealth health data (MBS, PBS) with national hospital inpatient and mortality data. This is entirely feasible now. Other high priority datasets for inclusion in a national linked data resource include aged care data, emergency department data, perinatal data, clinical quality registries, and disease and device registries.

At present only AIHW, the Australian Institute of Family Studies and the Australian Bureau of Statistics (ABS) are accredited to link Commonwealth data, and the latter has not yet established workable arrangements whereby health and medical researchers can access the detailed microdata that are needed for their work. The AIHW has a substantial backlog of health data linkage projects. Projects involving linkage of Commonwealth data with those from one or more States or Territories could efficiently be performed by the State-based data linkage units (such as the NSW Centre for Health Record Linkage), and accreditation of these units to link Commonwealth data would greatly facilitate more timely and greater access to linked health data by researchers. State-based linkage also has the potential to be more accurate as it is based on a Master Linkage Key compiled from multiple datasets that can then increase the accuracy of the linkage.

Furthermore, the ethics and governance of data linkage activities needs to be simplified to allow agile responses to emerging and contemporary research and policy questions. The

current system is underpinned by 'one-off' linkage and integration of data for specific purposes and slow, confusing and cumbersome processes to obtain data custodian and ethical approvals. For projects involving both Commonwealth and State-based data collections, approvals from multiple agencies and Human Research Ethics Committees are needed, requiring the researchers to address varying and sometimes contradictory queries, and generating unnecessary costs for all concerned. This means that rapid response to the key health policy issues of the day are not possible due to the lead time necessary to create such linkages. Relinking exactly the same data sets for different projects on a 'one-off' basis is neither efficient nor in any way necessary for the purposes of protecting privacy and confidentiality.

Private sector data

What constitutes high-value private sector datasets – characteristics and potential benefits?

High-value private sector data present similar potential benefits to high-value public-sector data. In health, these include data collected by private health insurers, private hospitals, general practices, and pathology and imaging laboratories, as well as data collected for research purposes, such as clinical trials, cohort studies and device registries. To be considered high-value they must also be complete and accurate, including with free-text fields capable of being mined to extract the relevant information. Guidelines/legislation on governance and access arrangements for consumers and stakeholders are needed to support the proliferation of private sector data collections.

Managing the costs

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

Infrastructure for sharing public sector data should be funded by governments, consistent with their current and emerging open data policies that recognise the value of sharing these data. For example, the Australian Government's Public Data Policy Statement[2] commits it to optimising the use and reuse of public data, releasing non-sensitive data as open by default and collaborating with researchers to extend the value of data. Similarly, the NSW Government Open Data Policy[3] requires agencies to manage data as a strategic asset to be open by default, discoverable and usable and free where appropriate.

Similarly, infrastructure for sharing private sector data should be funded by the data owners, recognising the potential value to their organisations of re-use of these data. Where private sector data owners are funded or regulated by the public sector, provisions should be put in place to require the supply of relevant data so it can be shared through the same infrastructure as public sector data. For example, as part of their licensing requirements, private hospitals are currently required to submit data monthly to the relevant State health department and this is included in the admitted patient data submitted to the AIHW for inclusion in the national inpatient data collection.

Marginal costs relating to specific research projects should be funded through research grants or the funders of commissioned or sponsored research. However, pricing models need to recognise the 'hand to mouth' nature of grant-funded research. Costs of data linkage, data supply and data access are currently unpredictable, vary widely between

different agencies that are providing similar services, and may not be covered by the funds initially budgeted in research grants (e.g. costs for using the SURE laboratory have risen steeply since its establishment).

Is availability of skilled labour (e.g. data scientists) an issue and is there a role for government in improving the skills base?

There is an acute shortage of data scientists in health, as in other sectors. Government could consider supporting one or more interdisciplinary institutes charged with training the next generation of data scientists. The Columbia University Data Science Institute (see: http://datascience.columbia.edu/columbia-data-science) presents a potential model. This was launched in 2012 as part of New York City's Applied Sciences NYC initiative, which seeks to dramatically expand capacity in the applied sciences to maintain the city's global competitiveness and create jobs in the wake of the 2008 Global Financial Crisis.

Trust and privacy issues:

- *What types of data pose the greatest concern for privacy protection?*
- *What weight should be given to privacy protection relative to the benefits of greater data availability and use?*
- *Are changes to the privacy-related policy framework needed and if so, what are they?*

Are security measures for public sector data too prescriptive?

How should the risks and consequences of data breaches be assessed and managed?

Adoption of a principled proportionate governance model (PPGM) based on clear guiding principles would help to overcome key impediments to using health records for research. Such models gauge potential risks associated with data uses and mitigations to those risks, including the potential public interest that is served by enabling research. They require a clear articulation of roles and responsibilities at all levels of decision-making and effective training for researchers and data custodians. Examples of PPGMs include those developed by the Scottish Health Informatics Program[4] and PopData BC[5].

Approved researcher accreditation could be considered as a risk mitigation strategy for research use of potentially sensitive microdata, such as linked health records. This would be preferable to crude 'one size fits all' approaches to disclosure control such as restricting the records able to be accessed (e.g. to a 10% sample) or recoding to limit the level of detail available (such as aggregating geographic areas), which reduce and sometimes negate the value of the data for research and discovery. Examples of approved researcher accreditation processes include those operated by the UK Office of National Statistics (https://www.ons.gov.uk/aboutus/whatwedo/statistics/requestingstatistics/approvedresearcheraccreditation) and UK Economic and Social Research Council (https://www.ukdataservice.ac.uk/get-data/how-to-access/accesssecurelab).

Data infrastructure

What are Australia's infrastructure needs to improve access and usability of public and private sector data?

From a health perspective, creation of a new national enduring and accessible resource of linked health data ('Health Big Data virtual laboratory') is the top priority. This would provide accredited researchers with secure remote access to a set of core linked health data resources, with a single streamlined approval process and without the requirement to interact with multiple data custodians, ethics committees and data linkage units. The first priority is to provide access to an anonymised national linked dataset comprising:

- hospital morbidity data
- MBS and PBS data
- mortality data

Over time, additional national datasets (e.g. cancer registrations, perinatal data, emergency department data, aged care data, electronic health record data and health-related data from other agencies) could be added.

Also a top priority is infrastructure (including human capital) to support large-scale partnership research programs focused on using big data to tackle major challenges for Australia. Existing research groups rely on unpredictable competitive grant funding and short-term contract research, and therefore struggle to achieve and maintain critical mass. This fragmented approach to funding means that resources are wasted on applying for research funding, processes for assessing these applications, and in the start-up phases of multiple small research projects. Large-scale, longer-term partnership research programs with government and industry would achieve better outcomes for Australia.

Examples

What are the benefits of greater availability of data to research and what benefits could researchers give back to government?

Whole of population data linkage is a powerful and cost-effective means of generating unbiased, long-term evidence about real world health care use and outcomes.

Australian governments invest more than $100 billion annually on healthcare, yet we have a relatively limited understanding of Australia's return on this investment. For example, the Medicare Benefits Scheme (Commonwealth spend approximately $21 billion annually) consists of some 6000 items, but fewer than 5% have been assessed for safety, effectiveness and cost-effectiveness against contemporary evidence. Even when medical treatments have undergone extensive pre-market evaluation in randomised controlled trials, like all of the items listed on the Pharmaceutical Benefits Scheme (Commonwealth spend approximately $9 billion annually), they are most often tested over relatively limited time frames, even if medicines are used for a lifetime, and in populations that do not represent the people using them in routine clinical care. Real-world studies underpinned by health data linkage are the only means of establishing the population benefit and risk profiles of these interventions.

Additionally, research using these data will help us to understand the determinants of disease risk, target therapies to those who will benefit most, compare the effectiveness of alternative interventions, and model the health and economic impacts of interventions and

policies. Recognising this potential, research that will lead to 'better models of care and services that improve outcomes, reduce disparities, increase efficiency and provide greater value' is listed first among the new national research priorities for health (http://www.science.gov.au/scienceGov/ScienceAndResearchPriorities/Pages/Health.aspx). Moreover, it is estimated that effective use of big data could also deliver reductions to national health care expenditure of around 8%, which would translate to more than $11 billion annually in Australia.[6]

Background: UNSW Australia's Centre for Big Data Research in Health (CBDRH)

The Centre for Big Data Research in Health (https://cbdrh.med.unsw.edu.au/) is a world-first research centre that aims to maximise the use of all possible sources of health big data in order to enhance the health and well being of Australians and the global community. All of its projects use public sector data, most use linked data and some involve linkage of public sector data with research data. The Centre's areas of research focus include health system performance, value and waste in health care, evaluating policies and programs using linked data, innovative analytic methods for health big data, increasing the use of evidence in policy and practice, cancer causes and care, drug safety, maternal and child health, chronic disease and multi-morbidity and Aboriginal health.

Examples of recent CBDRH research

How can the health system tackle disparities in Indigenous health?

The Indigenous Health Outcomes Patient Evaluation (IHOPE) project used advanced statistical modelling techniques, and "real-world" whole-of-population linked data from hospitals, to unpack the contributions of personal factors (such as socioeconomic status and comorbidities), hospital factors (such as hospital size and facilities) and geographic factors (such as remoteness) to disparities between Aboriginal and non-Aboriginal people in the incidence, management and outcomes of health conditions. This allowed identification of which strategies have the greatest potential for tackling these disparities. The issues investigated included incidence and outcomes of acute myocardial infarction, avoidable hospitalisation, otitis media in children, cataract surgery, road traffic injuries and unintentional injuries. The research found that crucial issues driving poor outcomes for Aboriginal people included high rates of comorbidities, low levels of private health insurance, use of smaller hospitals with fewer specialist services, and limited access to publically-funded services. IHOPE research has informed the planning of cardiac, ear and eye health services for Aboriginal people in NSW. IHOPE findings have been cited in 5 national and state policy documents, including the National Heart Foundation's Consensus Statement on management of Acute Coronary Syndrome in Aboriginal people.

IHOPE was honoured as one of NHMRC's '10 of the best' projects in 2015.

Did the ABC's Catalyst program change statin use in Australia?[7]

Population-based interrupted times series analysis of PBS data for more than 190,000 people who were dispensed statins was used to examine the impact of a two-part special edition of the Australian Broadcasting Corporation's science journalism program Catalyst (titled Heart of the matter), aired in October 2013, that was critical of HMG-CoA reductase inhibitors ('statins'). The analysis showed that following the Catalyst program, there was a

2.60% reduction in statin dispensing, equivalent to 14,005 fewer dispensings Australia-wide every week. In the week the Catalyst program aired, there was a 28.8% increase in discontinuation of statin use, which decayed by 9% per week.

An estimated 28,784 additional Australians ceased statin treatment. Up until 30 June 2014, there were 504,180 fewer dispensings of statins, and the authors estimated that this affected 60,897 people. The changes in statin use occurred despite warnings in the Catalyst program that its content should not be taken as medical advice, and public criticism of the program. This research showed the power of the media in influencing public opinion as well as the consequences when the media get it wrong.

The study won the 2015 MJA, MDA National Prize for Excellence in Medical Research.

What was the impact of the National Perinatal Depression Initiative on access to Medicare services for women at risk of perinatal mental illness?[8]

Retrospective cohort study using difference-in-difference analytical methods was used to quantify the impact of the National Perinatal Depression Initiative policies on MBS mental health usage by Australian women giving birth between 2006 and 2010. A random sample of women of reproductive age enrolled in Medicare who had not given birth where used as controls. The difference-in-difference analysis showed that while there was an overall increase in MBS mental health item use as a result of the National Perinatal Depression Initiative, this did not reach statistical significance. However, the National Perinatal Depression Initiative was found to significantly increase access in subpopulations of women, particularly those aged under 25 and over 34 years living in major cities. The researchers recommended that new mental health policy initiatives should incorporate a planned strategic approach to evaluation, which includes sufficient follow-up to assess their impact.

The Abbott-Turnbull Government abolished the National Perinatal Depression Initiative in 2015. The study was used to inform the Shorten Labor party's 2016 Election plan for Excellence in Perinatal Health.

http://www.100positivepolicies.org.au/national_plan_for_excellence_in_perinatal_mental_health_fact_sheet

Do higher doses of immunosuppression increase cancer risk in solid organ transplant recipients?[9] [10]

This cohort study linked the Australian Cancer Database, National Death Index, Australia and New Zealand Liver Transplant Registry, Australia and New Zealand Cardiothoracic Organ Transplant Registry and hospital medical record data to identify outcomes after liver, heart, and lung transplantation in Australia. The data uniquely captured changes over time in the type and dose of immunosuppression used to avoid organ rejection. With over two-decades of follow-up this study showed that higher mean daily doses of the immunosuppressive agent azathioprine more than doubled the risk of two of the most common cancers in transplant recipients, non-Hodgkin lymphoma and lip cancer. These findings have major implications for managing cancer risk in transplant recipients and other immunosuppressed populations, and for understanding the pathways to carcinogenesis for these malignancies.

Do multiple-birth children incur greater hospital costs than singleton-birth children and what is the role of Assisted Reproductive Technology?[11]

This study used linked data for all births in Western Australia over a ten-year period to hospital admission and death data during the first 5 years of life. A comprehensive economic and health services assessment of the frequency, duration, and cost of hospital admissions for singleton, twin, and higher-order multiple (HOM) children was performed, including an assessment of the contribution of assisted reproductive technology (ART) to the incidence and cost of multiple births. The study found that compared with singletons, twins and HOMs were 3.4 and 9.6 times, respectively, more likely to be stillborn and were 6.4 and 36.7 times, respectively, more likely to die during the neonatal period. Twins and HOMs were 18.7 and 525.1 times, respectively, more likely to be preterm, and 3.6 and 2.8 times, respectively, more likely to be small for gestational age. The mean hospital costs of a singleton, twin, and HOM child to age 5 years were $2730, $8993, and $24 411,respectively, with cost differences concentrated in the neonatal period and during the first year of life. Almost 15%of inpatient costs for multiple births could have been avoided if ART twins and HOMs had been born as singletons.

This study was published in *JAMA Pediatrics*, and accompanied by an independent Editorial calling for a change in the practice of transferring multiple embryos during IVF treatment. The study is in the top 5% of all research outputs scored by Altmetric.

[1] Jorm L. Routinely collected data as a strategic resource for research: priorities for methods and workforce. *Public Health Res Pract*. 2015;25(4):e2541540.

[2] Australian Government Public Data Policy Statement. Available at: https://www.dpmc.gov.au/sites/default/files/publications/aust_govt_public_data_policy_statement_1.pdf

[3] NSW Government Open Data Policy. Available at: https://www.finance.nsw.gov.au/ict/sites/default/files/resources/NSW_Government_Open_Data_Policy_2016.pdf

[4] Sethi N, Laurie GT. Delivering proportionate governance in the era of eHealth. Making linkage and privacy work together. *Med Law Int*. 2013 Jun; 13(2-3): 168–204.

[5] McGrail KM, Gutteridge K, Meagher NL. Building on Principles: The Case for Comprehensive, Proportionate Governance of Data Access. In Gkoulalas-Divanis A, Loukides G (eds.) *Medical Data Privacy Handbook*. Springer: Switzerland 2015, pp 737-764.

[6] Groves P, Kayyali B, Knott D, Van Kuiken S. *The 'big data' revolution in health care*. New York: McKinsey and Company Center for US Health System Reform, 2013.

[7] Schaffer AL, Buckley NA, Dobbins TA, Banks E, Pearson S. The crux of the matter: did the ABC's Catalyst program change statin use in Australia? *Med J Aust* 2015; 202 (11): 591-594.

[8] Chambers GM, Randall S, Hoang VP, Sullivan EA, Highet N, Croft M, Mihalopoulos C, Morgan VA, Reilly N, Austin MP. The National Perinatal Depression Initiative: An evaluation of access to general practitioners, psychologists and psychiatrists through the Medicare Benefits Schedule. *Aust NZ J Psychiatry* 2016; 50(3): 264-274.

[9] Na R, Laaksonen M, Grulich AE, McCaughan G, Keogh A, Meagher N, Vajdic CM. Iatrogenic immunosuppression and risk of non-Hodgkin lymphoma after liver, heart and lung transplantation: A population-based cohort study. *Brit J Haem* 2016 May 2 doi: 10.1111/bjh.14083.

[10] Na R, Laaksonen M, Grulich AE, McCaughan G, Keogh A, Meagher N, Vajdic CM. High azathioprine dose and lip cancer risk in liver, heart and lung transplant recipients: A population-based cohort study. *J Am Acad Dermatol* 2016; 74 (6): 1144-1152.

[11] Chambers GM, Hoang VP, Lee E, Hansen M, Sullivan EA, Bower C, Chapman MG. Hospital costs of multiple and singleton birth children during the first five years of life and the role of assisted reproductive technologies. *JAMA Pediatrics* 2014; 168(11):1045-1053 [Plus Editorial and Letter]

http://dx.doi.org/doi:10.1001/jamapediatrics.2014.1357

http://dx.doi.org/doi:10.1001/jamapediatrics.2014.1683 (Editorial)