In support of access to integrated public data for scientific research

Yohannes Kinfu, Tom Cochrane, Rachel Davey, Ivan Hanigan, Nasser Bagheri
Health Research Institute, University of Canberra

Scientific research needs to avoid false-positive or false-negative findings due to inadequate data, yet many public and private databases are either too limited for strong inferences to be made regarding complex issues or not readily available for researchers. Improved access to data (and especially linked databases) is an essential ingredient to resolving this to produce high quality research.

This brief submission supports expanding the 'data space' for academic research and provides specific examples of the opportunities that access to data, particularly integrated or linked data, creates for understanding the nation's most pressing socio-economic, demographic, and health challenges. Data linkage provides a longitudinal perspective to policy analysis and research, is integrative, efficient and cost-effective relative to other data collection methods, and enables analysis of certain relationship that could not otherwise be performed. We argue that the move towards better access to public data should go hand in hand with expanding data linkage programs.

Data linkage is not new in Australia. The Australian Bureau of Statistics (ABS), The Australian Institute of Health and Welfare (AIHW) and The Australian Institute of Family Studies (AIFS) are experienced integrating authorities and have carried out various valuable data integration exercises that led to findings of national significance and benefit. One exercise that deserves special mention is the integration of death records with the country's most recent censuses, which, for the first time, allowed the country to measure the gap in survival rates between Indigenous and non-Indigenous Australians with greater levels of accuracy (ABS 2006). Before the data integration exercise was undertaken the demography of Indigenous Australia remained elusive and varied widely depending on the models used (Kinfu and Taylor 2002). Commenting on these, Kinfu and Taylor (2002, 2005) wrote that:

> "In truth, presently it is impossible to determine the factors that contribute to non-demographic population growth [among Indigenous Australians]. Methodologically, this would require either a larger post-enumeration sample survey or the matching of unit records from one census to the next as well as with births and deaths registration data."

Indigenous disadvantage also persists in the areas of education, labour force outcome, incarceration, morbidity and hospital utilisations. The scientific and policy discourse in these areas are dominated by cross-sectional analyses that have a number of limitations. One area where improvement in understanding disadvantage in Australian communities could be made is by way of linking the various censuses with health facility, educational, or income and employment data. For example, in 2014 integrated Queensland Education and ABS census data has enabled the state to examine the effects of socio-economic factors on student achievement that could not otherwise be performed (ABS 2014). A second example is the integration of

NAPLAN data with the Longitudinal Study of Australian Children (LSAC) carried out by the AIFS. The integration of the Survey data will no doubt enable research on the effects of residential location and early life Socio-Economic Status (SES) on health and educational outcome that is not currently feasible. However, given the limited sample in the survey the potential for using the linked data for sub-national analysis will be limited and this is why we argue the LSAC data should be further linked with Census data to allow a refined spatial analysis of health and educational outcome of Australian children.

Linking the various agricultural surveys in the country with death records and environmental data would be very useful for the study of environmental threats to health. For example, such linked data will enable the country to have better insight on the relationship between farmer's suicide, agricultural income, climate change and environmental stress. Other approaches such as linking death records with tax data would enable the country to estimate the economic cost of premature mortality from certain conditions. Such information is vital for the country but cannot be undertaken at the moment because of lack of access to linked data.

Health challenges such as the emerging dominance of chronic diseases and lifestyle-related causes (Murray et al. 2015) and rising net costs of hospitalisations, which are evident from various health surveys and nationally reported data, provide a snapshot of national health status only but little insight into the causes, potential remedies and the impact of interventions on future public health. Integrating the country's various health surveys with hospital records would improve data quality; provide opportunity to examine cost of treatment and hospital utilisation patterns over a longer time span than currently elicited through self- report. The 45 Up study managed by the Sax Institute already links the data from its survey population with their hospital records but the fact that the survey is confined to a single jurisdiction, New South Wales, means that only a national health data linkage would be able to provide a complete picture for the country.

In short, while the call for increased access to public data from health, education and other sectors is a step in the right direction, it is only when these databases are linked with each other and with census and survey data and made available to a wide research community in a secure environment that the country would benefit most. Currently the number of linked datasets in the country are very limited and carried out as 'projects' rather than with the aim of building a broad database for understanding various aspects of the country's population. In addition, at present with the exception of some data from the AIFS individual researchers have limited access to existing linked databases as these data are often 'developed' only for internal purposes.

So what next?

There is a lot of experience out there that we can build on to move the agenda of data integration and access to integrated data in Australia further. Commenting on a recent workshop on linking Census, Survey and Administrative data in the US organised by the National Academy of Sciences, Engineering and Medicine in June 2016, Hout noted that the planned data integration exercise in the US, dubbed as "the American Opportunity Study" is the next big step and foundational for future

research in the US (National Academy of Sciences, Engineering, and Medicine). Participants of the workshop stressed that to make this study a success “We have to offer back to those who have the data the products and insight they could not otherwise get. We are not just bringing together data or assembling data but enhancing and passing back better, more usable information to those who have provided data to us.” (National Academy of Sciences, Engineering, and Medicine). This is one of the valuable lessons for Australia as it also plans to expand its own public data access efforts.

Another lesson from the US study that is also relevant to Australia is the reality of inter-jurisdictional differences in the type and quality of administrative data as well as the rules governing access to these data. While there is no simple answer to get around this issue, bringing all stakeholders together for a pre-implementation workshop such as the one conducted by the US National Academy of Sciences, Engineering and Medicine could be one way to cultivate confidence, trust and buy-in.

Data integration also has its own technical challenges; it requires statistical competence and computational infrastructure that allows the linkage to be undertaken with complete security and respect for privacy or commercial sensitivity. As we have mentioned earlier, Australia already has three accredited integrating authorities approved by the Cross Portfolio Data Integration Oversight Board. It is important that these are strengthened. However, some of these institutions should be encouraged (or even required) to share the linked data for research. Current experience is that this has been lagging behind relative to what is now common practice elsewhere (notably in the UK, Canada and the United States) and data generated through linkages still remain out of bounds to a wealth of talented researchers.

References

Australian Bureau of Statistics (ABS). 2014. Educational outcomes, experimental estimates, Queensland, 2011, ABS.

Australian Bureau of Statistics (ABS). 2006. Discussion Paper: Assessment of Methods for Developing Life Tables for Aboriginal and Torres Strait Islander Australians, ABS.

Kinfu, Y. Taylor, J. 2005. On the Components of Indigenous Population Change, Australian Geographer, Vol 36, No 2: 233-255.

Kinfu, Y. Taylor, J. 2002. Estimating the components of Indigenous population change, 1996–2001, Discussion Paper No 240/2002, Centre for Aboriginal Economic Policy Research, Australian National University. ISBN 0 7315 5615 1.

Murray, C et al. 2015. Global, regional, and national disability-adjusted life years (DALYs) for 306 diseases and injuries and healthy life expectancy (HALE) for 188 countries, 1990–2013: quantifying the epidemiological transition, **Volume 386, No. 10009**, p2145–2191, 28 November 2015.

National Academy of Sciences, Engineering, and Medicine. 2016. Using Linked Census, Survey, and Administrative Data to Assess the Longer-Term Effects of Policy: Proceedings of a Workshop—in Brief. Washington, DC: The National Academies Press. doi: 10.17226/23583.