PUBLIC HEALTH COMMUNICATION CENTRE AOTEAROA

Capitalising on NZ's linked data by increasing IT and researcher capacity: Opportunity is knocking

18 June 2018

Tony Blakely, Andrea Teng, Sheree Gibb, Nhung Nghiem, Barry Milne, Andrew Sporle , Gabrielle Davie, Nevil Pierce, Ruth Cunningham

A key strategic advantage for NZ and research is our national routinely-collected datasets. This can generate new knowledge in academia, service delivery and policy. Conversely, NZ has some key barriers to overcome to make the best use of that data – most importantly, data systems infrastructure and research capacity. In this blog we consider these opportunities and barriers. We believe we are at a moment in time when a major centralized investment is required that will return dividends to NZ citizens and academics through better policy making and new knowledge discovery.

NZ has a key natural advantage in research: our national datasets. NZ is also a small country with strong links between policy makers and researchers, and a collaborative environment, strengths we can build on.

Consider the health data. All mortality and hospitalisation events, cancer registration, publicly funded community pharmacy dispensings, lab tests, and so on, can be linked anonymously through the National Health Index number (unique patient identifier). Moreover, they are now linked to other administrative datasets and survey data in the Statistics NZ Integrated Data Infrastructure (SNZ IDI).

NZ's data systems are world leaders in many ways, with the IDI gaining a range of international interest from other governments wishing to implement similar systems. The IDI contains linked data about people and households, from government agencies, Stats NZ surveys, and non-government organisations. This includes data about health, income and work, education and training, justice, corrections, social development, immigration, housing, people and communities and more. The IDI provides a single, centralized system governed and managed by StatsNZ. In Australia, there are multiple state and federal based data collations that do not routinely share data, with tortuous application processes.

Here is the vision of the Virtual Health Information Network:

To create and sustain an environment that captures value from linking health data collections, through world leading health research, policy development and service planning.

In more immediate terms, the vision is about scaling up capacity (IT infrastructure, human resource) – because at the moment the current capacity is severely limiting the benefits we can get from our data resources.

So what are the barriers and gaps?

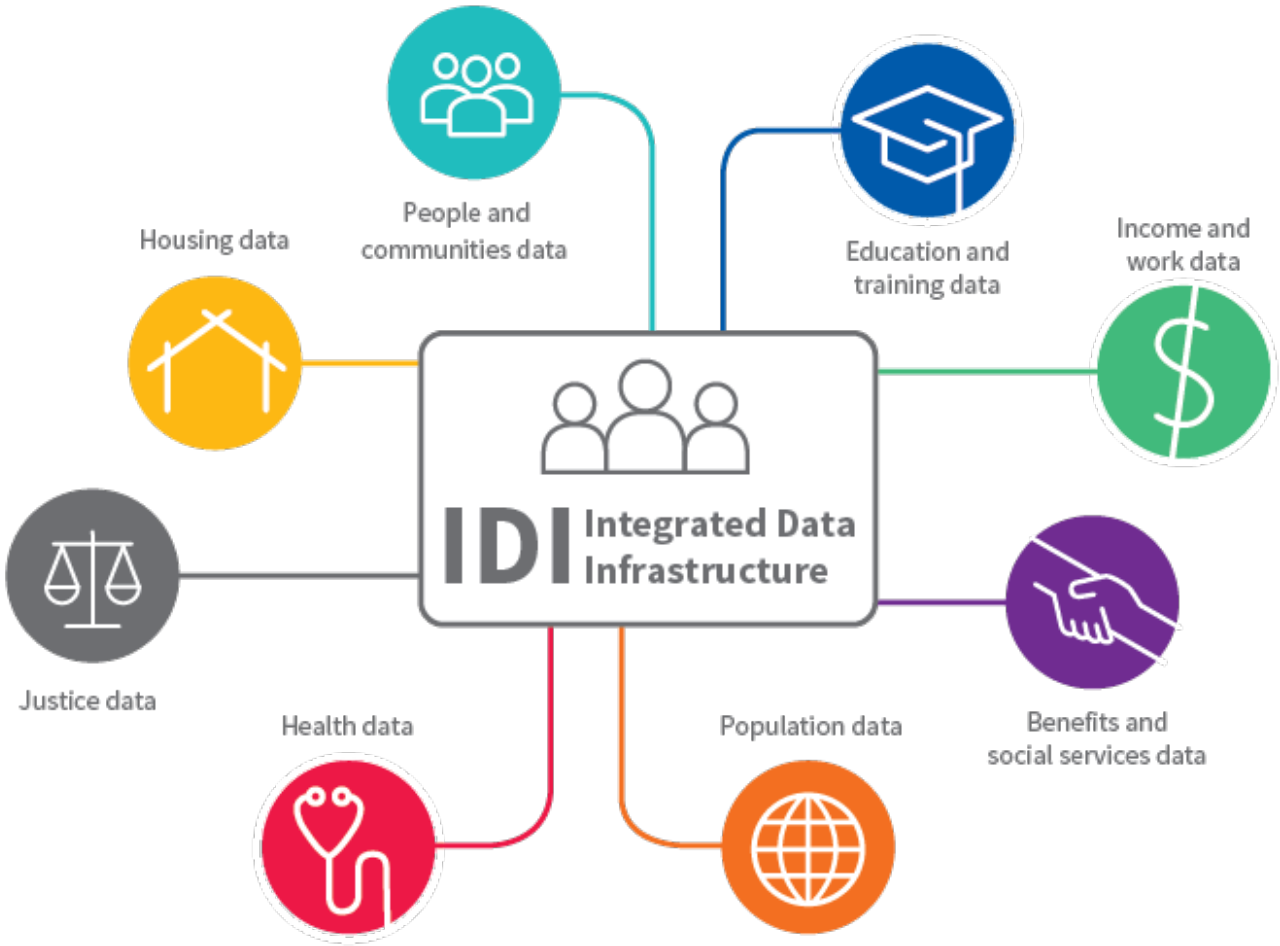
First, **researcher capacity** for dealing with:

- missing, messy and mis-measured data
- longitudinal analyses (often researchers analyse longitudinal data as though it is repeated cross-sectional data, failing to capitalize on within individual change)
- contemporary causal methods that much better address sources of error in the context of observational data (for example, causal mediation analyses [1 2] and a NZ example [3]), and analyses of time varying exposures and confounders [4 5])
- big data and the opportunities with machine learning [6-8] (apparently, there is only one machine learning project in the SNZ IDI to this point).

This researcher capacity limit has a number of causes, from being a small country, to lack of advanced training opportunities (especially post-doc), lack of support for quantitative skills in the health and social sciences, absence of a critical mass in advanced statistical methods and computer science in NZ, and so on. If NZ is to realise a dividend from its world-leading data resources then we need to seriously upskill the workforce. Two of us (TB, AS) are facilitating international researchers visiting to share skills, but that should be viewed as part of capacity building – not the answer.

Second, we need better **IT and data systems**. Accolades first – Stats NZ has done an amazing job expanding nationally linked datasets in one place, namely the StatsNZ IDI.

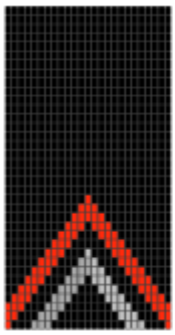
However, we also need to front up to the reality – the IT infrastructure in the StatsNZ IDI is not designed for the high intensity applications that will realise the value of the data resource that NZ should be leading the world in. The best data needs the best methods, but NZ's current official data IT systems are not capable of supporting the recent advances in big data analytic methods. For example, one of us (NN) is unable to make reasonably basic machine learning algorithms work in the SNZ IDI due to memory and server issues. There are new efficient tools for accessing and analysing big data securely (eg N1 Analytics); in NZ we are using old technology to try to gain new insights.



So, what are the solutions? Many. They do involve investing in resource – both people and IT infrastructure. We sense that this may be about to change – for the better. The health-related National Science Challenges are laying the groundwork for a step up in big data and life course research, the VHIN is lobbying behind the scenes, and we know from numerous discussions with many colleagues that there is a reasonably common view of the problems and opportunities across Government and academic agencies (albeit not officially stated). Moreover, MBIE is signalling new funds within the Strategic Science Investment Fund in 2019.

Stats NZ recognises these barriers and has an expansion programme known as IDI2 looking at how they can increase the use and improve the usability of the IDI. As users have seen what is possible with integrated data, user demands have changed and become more complex than what the IDI was originally set up for. They are also investigating what is the core purpose and potential for integrated data so we can have the necessary tools now and in the future that best help users gain insights to improve New Zealanders' lives.

It is critical to acknowledge, and facilitate, good governance systems for big data. Current measures include Stats NZ balancing the benefit insights can bring with protecting privacy and security. Under their privacy and security requirements, all data has had identifying information such as names and dates of birth removed. Only vetted and approved researchers can access selected, deidentified datasets for specific projects. Such research must have a public good focus. Users can only access the IDI in secure research facilities. Thinking ahead, good work is in progress with the Data Futures Partnership this needs continuing support, realising the shape and boundaries of usable data are rapidly changing. None of the above opportunities to captialize on NZ integrated data are going to happen if NZ citizens do not judge that their data is being appropriately protected and respected. Conversely, NZ citizens expect good use to be made of their data to justify its collection. Social license needs to be monitored, maintained and enhanced. In particular, Māori data sovereignty is important. Ensuring appropriate use of data resources requires a single ethical oversight system that is inclusive of community values and transparent in its processes.



TE MANA RARAUNGA

Māori Data Sovereignty Network

New Zealand research largely works on a semi-competitive model with loose collaboration between academics and analysts getting on with good research – be they in universities, government agencies or elsewhere. Till the soil, let many flowers bloom, and see what thrives. But occasionally an opportunity or need arises that requires large critical mass and – for a period of time at least – a kick-start of serious resources. Australia has achieved this with Data61 – a national partnership between the government's CSIRO and multiple research institutions, focussed solely on advances in data analytics with broad research and policy application. And several Australian universities are making major investments. We believe we are at a moment where NZ needs a similar approach, not from just a health research point of view, but actually for our wider science system. We have an opportunity now to transform our world leading data into a strategic resource that works for research, policy and the wider community to inform our development at local and national levels. A collaborative centre of 20-40 analysts and academics, with a sound business plan, international collaborations and strong community engagement, might be one way to go. But it is not going to happen without planning.

An effective collaborative centre will need to position itself so it can answer the research questions that are meaningful to the lives of NZers, and it will need a strong home in academia – not just a Government agency (ie to complement Government agencies like the Social Investment Agency). Can NZ achieve this? It will require a Minister, Government Agency Chief Executive or Vice Chancellor somewhere to 'make it happen'. The good news is that the awareness within the sector for such a bold – but necessary – initiative seems to be coalescing.

Authors: Tony Blakely, Andrea Teng, Sheree Gibb, Nhung Nghiem, Barry Milne, Andrew Sporle, Gabrielle Davie, Nevil Pierce, Ruth Cunningham, and on behalf of the Virtual Health

Information Network

References

1. VanderWeele TJ. Explanation in Causal Inference: Methods for Mediation and Interaction: Oxford University Press 2015.
2. VanderWeele TJ. Mediation Analysis: A Practitioner's Guide. *Annu Rev Public Health* 2016;37:17-32. doi: 10.1146/annurev-publhealth-032315-021402 [published Online First: 2015/12/15]
3. Blakely T, Disney G, Valeri L, et al. Socioeconomic and Tobacco Mediation of Ethnic Inequalities in Mortality over Time: Repeated Census-mortality Cohort Studies, 1981 to 2011. *Epidemiology* 2018;29(4):506-16. doi: 10.1097/EDE.0000000000000842 [published Online First: 2018/04/12]
4. Cole SR, Hernan MA, Robins JM, et al. Effect of highly active antiretroviral therapy on time to acquired immunodeficiency syndrome or death using marginal structural models. *American Journal of Epidemiology* 2003;158(7):687-94.
5. Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology* 2000;11(5):550-60.
6. Mooney SJ, Pejaver V. Big Data in Public Health: Terminology, Machine Learning, and Privacy. *Annu Rev Public Health* 2018;39:95-112. doi: 10.1146/annurev-publhealth-040617-014208 [published Online First: 2017/12/21]
7. Naimi AI, Balzer LB. Stacked generalization: an introduction to super learning. *European Journal of Epidemiology* 2018;33(5):459-64. doi: 10.1007/s10654-018-0390-z
8. Keil AP, Edwards JK. You are smarter than you think: (super) machine learning in context. *Eur J Epidemiol* 2018 doi: 10.1007/s10654-018-0405-9 [published Online First: 2018/05/11]

Public Health Expert Briefing (ISSN 2816-1203)

Source URL:
https://www.phcc.org.nz/briefing/capitalising-nzs-linked-data-increasing-it-and-researcher-capacity-opportunity-knocking