International Journal of Epidemiology, 2019, 677–677e
doi: 10.1093/ije/dyz014
Advance Access Publication Date: 21 February 2019
Data Resource Profile

Data Resource Profile

Data Resource Profile: The New Zealand Integrated Data Infrastructure (IDI)

Barry J Milne,[1]* June Atkinson,[2] Tony Blakely,[2] Hilary Day,[2] Jeroen Douwes,[3] Sheree Gibb,[2] Meisha Nicolson,[4] Nichola Shackleton,[1] Andrew Sporle[5] and Andrea Teng[2]

[1]Centre of Methods and Policy Application in the Social Sciences (COMPASS), University of Auckland, Auckland, New Zealand, [2]Department of Public Health, University of Otago, Wellington, New Zealand, [3]Centre for Public Health Research, Massey University, Wellington, New Zealand, [4]Data and Digital, Ministry of Health, Wellington, New Zealand and [5]Department of Statistics, University of Auckland, Auckland, New Zealand

*Corresponding author. Centre of Methods and Policy Application in the Social Sciences (COMPASS), University of Auckland, 20 Wynyard Street, Auckland 1010, New Zealand. E-mail: b.milne@auckland.ac.nz

Editorial decision 18 January 2019; Editorial decision 18 January 2019; Accepted 13 February 2019

Data resource basics

The Integrated Data Infrastructure (IDI) is a collection of New Zealand whole-population administrative data sources from government agencies, the 2013 Census and several questionnaire-based social and socioeconomic surveys from samples of the population. The IDI allows whole-population analysis across different sectors of government (e.g. health, social services, education).

Data are available and linkable at the individual level for an 'ever resident' New Zealand population, including 'people born in New Zealand, permanent residents, people with visas that allow them to reside, work, or study in New Zealand (including international students and temporary workers), and those who live and work … [in New Zealand] without requiring a formal visa'.[1] Data provide a longitudinal record of events (e.g. hospitalizations, pharmaceuticals dispensed) over time, with different datasets covering different periods (e.g. hospitalizations have been captured since 1988 but maternity data only since 2002) (see Table 1). As of September 2018, the IDI holds 166 billion pieces of information.[2]

The IDI was first established in 2011, but has been continually updated with the addition of new records, new data tables and new fields within tables. The information presented in this profile is current as at September 2018.

Data collected

As New Zealand's national statistics office, Statistics New Zealand (Stats NZ) has a data and information leadership role throughout the public sector, and is funded through 'Vote Statistics'.[3] This role includes overseeing the IDI on behalf of the system. Data are sourced from a variety of public sector organizations (e.g. Ministry of Health, Ministry of Justice) and non-government organizations (e.g. Auckland City Mission), who are responsible for maintaining and providing up-to-date data tables to Stats NZ for inclusion in the IDI.

IDI data are stored in separate data tables in an SQL database (see Table 1 for an overview of data available in the IDI). Depending on the table, data may be at the individual, household, provider or small-area geographical level. Personal identifiers (i.e. names, addresses, government agency ID) have been removed to protect privacy and confidentiality; instead, unique but anonymous identifiers identify individuals and households. The main unique identifier for individuals (snz_uid) enables the same individual to be linked across different datasets; household- or area-level tables typically link to individuals using concordance tables linking household- or area-level identifiers to the snz_uid.

Table 1. Data available (start date) in the New Zealand Integrated Data Infrastructure (IDI)

Health	Social Services	Education	Justice/regulatory	Geographical and housing	Economic and socioeconomic	People and communities
Hospitalizations (public, 1988-; private, 2001-)	Benefits (1990-)	Tertiary education (1994-)	Vehicle registrations and driver licensing (1954-)	Social housing register (1980-)	Business register (1999-)	Births, deaths and marriages (1848-)
Mortality (1988-)	Child social services (1991-)	Training programmes (2001-)	Court charges (1992-)	Travel and migration (1997-)	Tax and income (1999-)	Auckland City Mission (vulnerable/ homeless population; 1996-)
Cancer registration (1995-)	Student loans and allowances (1992-)	Secondary education (2004-)	Sentencing and remand (1998-)	Geographical location (small area level; 2000-)	Longitudinal household panel survey (2002-10)	Longitudinal immigration survey (2005-)
Disabilities (1998-)	Injury claims (1994-)	Primary education (2007-)	Crime offenders (2009-)	Tenancy bonds (2000-)	Household labour force survey (2006-)	General social survey (2008-)
General Medical Service claims (fee-for-service primary health care; 2002-)	In work tax credits (2003-)	Early childhood education (2008-)	Victims of crime (2014-)		Household economic survey (2006-)	Migrant survey (2012)
Laboratory tests (2003-)	Youth services (2004-)				Income survey (2006-)	New Zealand Census 2013
Maternity care (2003-)	Home visiting intervention (2008-)					Māori General Social Survey (Te Kupenga; 2013)
Primary health care enrolments (2003-)						
Pharmaceutical dispensing (2005-)						
Immunizations (2005-)						
Chronic conditions (2007-)						
Outpatient attendance (2007-)						
Community mental health treatment (2008-)						
B4 School Check (health screen; 2011-)						
Disability survey (2013)						

The IDI is updated ('refreshed') up to four times per year, which includes addition of new datasets and updates of existing data. For each new refresh, datasets are linked to the IDI 'spine' using probabilistic linkage (the Fellegi-Sunter method).[4] The IDI spine aims to include all people who have ever been a resident in New Zealand, and it is constructed for each new refresh by linking together tax records since 1999, New Zealand birth records from 1920, and long-term visas from 1997 (currently about 10 million individuals).[1,2]

Linkage rates and linkage quality (e.g. false-positive error rates) differ for each dataset linkage to the spine. For health records linked to the spine, the linkage rate was estimated in September 2018 to be 85% (88% among New Zealand residents) and the false-positive rate is estimated to be 0.8%.[5,6] There are differences in linkage rates by sex (males are about 2% higher), ethnicity (lower for Pacific and Asian New Zealanders compared with Māori and European New Zealanders), and age-by-ethnicity (declines for European New Zealanders older than 85 years, Māori New Zealanders older than 65 years and Pacific and Asian New Zealanders older than 35 years).[6]

Data resource use

Under the 'Five Safes' framework followed by Stats NZ, IDI data can only be used for public-good research purposes.[7] IDI data cannot be used for individual case management or for regulatory purposes.

A searchable database of all IDI research projects is available via the Stats NZ website.[8] It is a condition of access that all projects and lead researcher details are published on this website, to ensure transparency about how the IDI is being used.

Broadly speaking, the IDI data can be used for four types of research: descriptive, analytical, methodological and evaluation of policies and interventions. Examples of each of these are provided below.

Descriptive research

Shackleton *et al.*[9] used repeated cross-sections of B4 School Check data in the IDI to identify decreasing trends in pre-school obesity in New Zealand from 2010 to 2016. McLeod *et al.*[10] used IDI data to identify and characterize adolescents who experience poor health and other outcomes.

Analytical research

IDI data can be used to define population cohorts which can be assessed over time.[11] For example, Dixon[12] defined cohorts of sufferers of chronic disease and assessed short- to mid-term effects on work and income; Berry *et al.*[13] assessed long-term outcomes for a cohort of very pre-term babies; Donovan *et al.*[14] used a cohort approach to show that living near green and more biodiverse vegetation lowered the risk of asthma; Davie and Lilley[15] assessed the financial impacts for an older cohort who have experienced injury; and Teng *et al.*[16] examined a cohort of earthquake survivors and showed that cardiovascular disease admission rates after the Canterbury earthquakes were associated with area-level residential damage. IDI data have also been used to assess socioeconomic and ethnic inequalities (e.g. in pre-school oral health[17] and in immunization[18]).

Methodological research

Methodological research using IDI data is an under-explored area, especially in regard to data quality and linkage bias and what can be done to overcome these issues. However, Zhao *et al.*[11] and Stats NZ, through their work on 'Census transformation',[19,20] have conducted methodological investigations regarding the ability of administrative data from the IDI to produce population estimates as accurately as the New Zealand Census.

Policy and intervention research

The IDI lends itself to evaluation work because researchers are able to assess the long-term impact of life events, policies or interventions for individuals. For example, Vaithianathan *et al.*[21] compared health and other outcomes for children and mothers who received the 'Family Start' intervention, a home visiting programme for at-risk and low-income mothers, against a propensity-scored-matched control group.

Strengths and weaknesses

Strengths

The IDI has several analytical advantages. First, the IDI links health data to data from various government sectors. This adds enormous value to already existing health data, in determining both the drivers of health and the consequences of ill health.

Second, the IDI allows for whole-population data analysis; few other countries have this ability. Many datasets in the IDI have national coverage and contain service use data for the whole population of New Zealand. The large sample sizes allow for analysis of small groups and rare events in ways that are not possible in projects that are dependent on primary collection of new data. Further, methods have been developed to accurately estimate the population under investigation at specific points in time (e.g. for use as a denominator), taking account of border movements, births and deaths.[11]

Third, the IDI's use of administrative data eliminates the risk of recall bias, which is a problem if data collection relies on self-reports of service use (e.g. hospitalization or pharmaceutical dispensing).

Fourth, the IDI can be used to analyse other information about study participants for whom detailed information was obtained in field studies at baseline (e.g. birth cohort studies), thus allowing longitudinal analyses of health outcomes identified through the IDI. In both cases, this requires loading data from field studies to the IDI. This is done by Stats NZ upon request (with anonymity being preserved throughout the process), so long as appropriate permissions (e.g. consent) are in place. However, research fully capitalizing on the longitudinal nature of the data (e.g. estimates of cumulative exposure or exposure patterns using marginal structural models or other G-methods) is rare to this point—an opportunity that should be utilized.[22]

Fifth, there are analytical advantages that have been touched on above: the ability to define and analyse long-term, system-wide trends for cohorts, and the ability to evaluate interventions.

Weaknesses

One key weakness is data quality. Data have been collected by a number of agencies, and by multiple people within multiple divisions within those agencies. Also, for the most part, data have been collected for operational, monitoring or accounting purposes, rather than for research purposes. As such, quality of data is variable across different agencies and across different data tables. However, quality of data is improving over time. For example, the pre-school health screen (B4 School Check) had incomplete coverage in early years, but more than 90% coverage from mid 2012. Also, pharmaceutical dispensing was not always recorded against a specific patient in the early years of the collection, but was recorded for more than 97% of all pharmaceuticals dispensed from mid 2007.

A second weakness is incomplete data documentation and metadata. Because the data have only recently been made available for wider research use, comprehensive documentation has lagged behind data availability. Furthermore, documentation about measures and constructs is stored at the agency level, with no central database of the measures contained in the IDI.

A third weakness is incorrect linkage. Linkage is probabilistic and, although the estimated false-positive rates are low (typically <2%),[5] incorrect linkage will cause errors in analyses of exposure–outcome associations—via selection bias (i.e. systematically missing observations), measurement error (e.g. incorrect variables from the wrong linked person assumed to apply to the index person) or confounding (e.g. missing confounders or mis-measured confounders). Rates of missing links (i.e. false-negatives) are unknown, and these may introduce similar biases. Additionally, as datasets are linked via a central spine, links between datasets external to the spine necessarily involve multiple linkage steps, each with potential error. For example, links between health and benefit data involve links between health data and the spine, as well as links between benefit data and the spine (i.e. two links and two possibilities for linkage error). Furthermore, and as noted above, health data for Pacific and Asian New Zealanders and older Māori New Zealanders are linked to the IDI spine at a lower rate, suggesting inconsistent coverage across population subgroups.[6]

A fourth weakness is that there will be no administrative data for individuals who have not accessed government agencies and services. This means, for example, that the IDI will not include individuals with health problems who do not access services for those health problems; and that only individuals charged or convicted for crimes will be included, which will be a (biased) subset of individuals committing crimes. Further, when a research question is reliant on data from multiple agencies, the proportion of individuals with data across all datasets may be low (i.e. if the proportion of individuals accessing services from all of the agencies is low). There is also the problem that a non-event (e.g. no hospital admission) is indistinguishable from a non-linked event (e.g. a hospital admission that was unable to be linked).

A fifth weakness is current limitations with Stats NZ IT infrastructure which limit the ability to use machine learning and other computationally intense methods.

Finally, researchers need to be aware that different datasets cover different periods (summarized in Table 1). As roughly two-thirds of data collections began in the year 2000 or later, this limits some longitudinal investigations. For example, long-term follow-up of children from the B4 School Check health screen is not yet feasible, as that collection only began in 2008.

Data resource access

Access to the IDI is by application to Stats NZ [https://www.stats.govt.nz/integrated-data/access-microdata-in-the-data-lab/]. Stats NZ applies the 'Five Safes' framework to statistical disclosure control—Safe People, Safe Projects, Safe Settings, Safe Data and Safe Output[7]—and researchers wishing to access the IDI are required to work within this framework.

Researchers applying to use IDI data are vetted by Stats NZ (Safe People). Researchers must supply a curriculum vitae and the names of two referees, and Stats NZ uses this information to assess whether the researcher is a bona fide researcher, belongs to a bona fide research institution, has a history of trustworthy data use and has an ability to analyse large datasets. Projects are also vetted (Safe Projects) to ensure they are public-good research, which analyses and reports on groups of people rather than individuals.

Under the Statistics Act 1975, Stats NZ are legally required to protect the privacy and confidentiality of the people and businesses they hold information about. Once a project is approved, first-time researchers participate in a one-hour training session on the confidentiality requirements of using IDI data. Researchers must sign a Confidentiality and Secrecy Agreement before accessing the data.

Data access for the project is then provided through a secure 'Data Lab' environment (Safe Settings). Researchers access data through a protected virtual environment and only in secure research facilities on computers that can access the IDI server, but nothing else (i.e. computer hard and soft drives cannot be accessed, and there is no access to the worldwide web). As such, IDI data are never sent to researchers, but instead access is granted to analyse data within the Data Lab environment. Data Lab facilities exist in Stats NZ offices and some government departments, universities and research agencies throughout New Zealand. Researchers from outside New Zealand can and do work on IDI projects but need to either travel to New Zealand to do so or collaborate with researchers able to undertake analyses in a New Zealand Data Lab.

Researchers access their project(s) through a login procedure that 'unlocks' only the data for which access has been granted; all identifiers have been removed from these data (Safe Data). Further value can be added by a researcher applying to link new external datasets into the IDI. An agreement between the data owners and Stats NZ is made to facilitate the new data linkage.

The statistical and programming packages SAS, STATA, R and Python are available to use for data analysis, as well as SQL for database management. Researchers can request results (but not individual-level data) to be released from the Data Lab, after they have applied confidentialization procedures to these results to ensure neither individuals nor attributes of individuals can be identified (Safe Output). Released output must also be accompanied by a standard disclaimer, indicating (among other things) that access to the data is provided under the Statistics Act (1975) and that 'careful consideration has been given to the privacy, security and confidentiality issues associated with using administrative and survey data in the IDI'.[23]

Support for IDI users is available through a Virtual Health Information Network [www.vhin.co.nz], which offers online guides, a discussion forum for users, a shared code repository and courses for users getting started in the IDI.

Ethics

The New Zealand Ministry of Health requires researchers to apply to the New Zealand Health and Disability Ethics Committee if they are requesting access to health data in the IDI [https://ethics.health.govt.nz/]. Stats NZ require that a researcher's organization supports the research proposal, but do not require institutional ethical review of IDI research projects. However, institutional ethical review may be a requirement for universities, government agencies and other institutions, as well as for journals and funders.

IDI in a nutshell

- The New Zealand Integrated Data Infrastructure (IDI) was set up to allow whole-population analysis across different sectors of government (e.g. health, social services, education).
- Established in 2011, the IDI has been continually updated with new records and data tables. Data are available and linkable for an 'ever resident' New Zealand population (as at September 2018, about 10 million individuals).
- Data are secondary administrative data. Data fields may be at the individual, household, provider or small-area geographical level.
- Data capture individuals' interactions with government agencies in the areas of health, social services, education, justice, geography, housing and economics (tax and income). Data from the 2013 Census and several questionnaire-based social and socioeconomic surveys from samples of the population are also included.
- Access to the IDI is by application to Stats NZ [https://www.stats.govt.nz/integrated-data/access-microdata-in-the-data-lab/]. Data access is through a secure 'Data Lab' environment, which exists in Stats NZ offices and some government departments, universities and research agencies throughout New Zealand. Researchers from outside New Zealand can work on IDI projects but need to travel to New Zealand to do so.

Acknowledgement

The authors would like to acknowledge Lucy Ashby from Stats NZ for useful comments on drafts of the manuscript.

Conflict of interest: None declared.

References

1. Black A. *The IDI Prototype Spine's Creation and Coverage*. (Statistics New Zealand Working Paper No 16–03). Wellington: Statistics New Zealand, 2016.
2. http://www.stats.govt.nz/browse_for_stats/snapshots-of-nz/integrated-data-infrastructure.aspx (26 October 2018, date last accessed).
3. New Zealand House of Representatives. *2017/18 Estimates for Vote Statistics. Report of the Government Administration Committee*. 2017. https://www.parliament.nz/resource/en-NZ/SCR_74572/457b275edea9d72dfb088e5fb256e2f751c08135 (26 October 2018, date last accessed).
4. Fellegi I, Sunter A. A theory of record linkage. *J Am Stat Assoc* 1969;**64**:1183–210.
5. Stats NZ. Integrated Data Infrastructure (IDI) Refresh: SM Monitoring Report. Statistical Methods, September 2018 Refresh. 2018. www.stats.govt.nz (26 October 2018, date last accessed).
6. Stats NZ. *March 2017 Integrated Data Infrastructure (IDI) Refresh: Linking Project Summary*. 2017. www.stats.govt.nz.
7. Desai T, Ritchie F, Welpton R. *Five Safes: Designing Data Access for Research*. Economics Working Paper Series 1601. Bristol, UK: University of the West of England, 2016.
8. https://cdm20045.contentdm.oclc.org/digital/collection/p20045coll17 (26 October 2018, date last accessed).
9. Shackleton N, Milne BJ, Audas R *et al*. Improving rates of overweight, obesity, and extreme obesity in New Zealand 4-year-old children in 2010-2016. *Pediatr Obes* 2018;**13**:766–77.
10. McLeod K, Ball C, Tumen S, Crichton S. *Using Integrated Administrative Data to Identify Youth Who Are at Risk of Poor Outcomes as Adults*. New Zealand Treasury Working Paper 15/02. Wellington: New Zealand Treasury, 2015.
11. Zhao J, Gibb S, Jackson R, Mehta S, Exeter DJ. Constructing whole of population cohorts for health and social research using the New Zealand Integrated Data Infrastructure. *Aust N Z J Public Health* 2018;**42**:382–88.
12. Dixon S. *The Employment and Income Effects of Eight Chronic and Acute Health Conditions*. New Zealand Treasury Working Paper 15/15. Wellington: New Zealand Treasury, 2015.
13. Berry MJ, Foster T, Rowe K, Robertson O, Robson R, Pierse N. Gestational age, health, and educational outcomes in adolescents. *Pediatrics* 2018;**142**:e20181016.
14. Donovan GH, Gatziolis D, Longley I, Douwes J. Vegetation diversity protects against childhood asthma: results from a large New Zealand birth cohort. *Nat Plants* 2018;**4**:358–64.
15. Davie G, Lilley R. Financial impact of injury in older workers: use of a national retrospective e-cohort to compare income patterns over 3 years in a universal injury compensation scheme. *BMJ Open* 2018;**8**:e018995.
16. Teng AM, Blakely T, Ivory V, Kingham S, Cameron V. Living in areas with different levels of earthquake damage and association with risk of cardiovascular disease: a cohort-linkage study. *Lancet Planet Health* 2017;**1**:e242–53.
17. Shackleton N, Broadbent JM, Thornley S, Milne BJ, Crengle S, Exeter DJ. Inequalities in dental caries experience among 4-year-old New Zealand children. *Community Dent Oral Epidemiol* 2018;**46**:288–96.
18. Charania NA, Paynter J, Lee AC, Watson DG, Turner NM. Exploring immunisation inequities among migrant and refugee children in New Zealand. *Hum Vaccin Immunother* 2018, Jul 19.–8. doi:10.1080/21645515.2018.1496769.
19. Stats NZ. Experimental Population Estimates from Linked Administrative Data: 2017 Release. 2017. http://archive.stats.govt.nz/methods/research-papers/topss/exp-pop-est-from-link-admin-data-2017.aspx (26 October 2018, date last accessed).
20. Suei S. *Comparing Income Information From Census and Administrative Sources*. 2016. http://archive.stats.govt.nz/methods/research-papers/topss/comp-income-info-census-idi.aspx (26 October 2018, date last accessed).
21. Vaithianathan R, Wilson M, Maloney T, Baird S. *The Impact of the Family Start Home Visiting Programme on Outcomes for Mothers and Children*. Wellington: Ministry of Social Development, 2016. ismatch]
22. Naimi AI, Cole SR, Kennedy EH. An introduction to G methods. *Int J Epidemiol* 2017;**46**:756–62.
23. Statistics New Zealand. *Microdata Output Guide*. 4th edn. Wellington: Statistics New Zealand, 2016.