Ten Years of the Integrated Data Infrastructure (IDI): What have we learned and what are the unresolved issues?

Barry Milne

COMPASS Seminar Series
3 May 2022
1. What is the IDI?

An integrated database containing confidentialised microdata about people and households

- Linked at the person-level (for most data tables)
  - Linkable, through tax data, to ‘Longitudinal business database (LBD)’

- ‘Ever resident’ population

- Longitudinal

- Intergenerational
1. What is the IDI?
History of the IDI

- 2011 IDI prototype
  - Labour market focused to allow linking of individual, household, and business-level data, with the following components
    - iLEED (integrated Longitudinal Employment and Education Data)
    - Migration and movements data
    - Longitudinal Business Database (LBD) data
  - Privacy impact assessment (Feb 2012)
- 2013 Cabinet agree to expand IDI – Better Public Services
  - Seven additional privacy impact assessments (June 2013 – July 2016) as additional datasets (ACC, Education, Justice, Census, health) were added
  - Overarching privacy impact assessments July 2017
- 2020-2021 Covid-19 data (wage subsidies, Immunisation Register)
- New data additions undergo a ‘Data Ethics and Privacy Assessment’

The five safes

We use a ‘five safes’ framework to ensure that we provide access to data only if all of these five conditions are met.

The ‘five safes’ framework for the IDI

1. Safe people
2. Safe projects
3. Safe settings
4. Safe data
5. Safe output
Data access

- Application deadlines every 6 weeks
  - Describe project (public good research)
  - List researchers
    - All need to be confidentiality trained
  - List data requested
    - Only the tables you are approved to use will be ‘unlocked’

Ngā Tikanga Paihere framework. Need to demonstrate:
  - Experience working with the population(s) of interest
  - Populations of interest support the research, and have been (and will continue to be) consulted about the research
  - The value of the research to the populations of interest
  - How potential risks (if any) to the populations of interest will be managed/mitigated
## 2. What have we learned?

### i. The IDI is a popular research tool

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Projects using integrated data</td>
<td>7</td>
<td>14</td>
<td>22</td>
<td>36</td>
<td>62</td>
<td>64</td>
<td>96</td>
<td>97</td>
<td>78</td>
<td>81</td>
<td>20</td>
</tr>
<tr>
<td>Projects not using integrated data</td>
<td>13</td>
<td>12</td>
<td>18</td>
<td>28</td>
<td>8</td>
<td>11</td>
<td>11</td>
<td>9</td>
<td>4</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>Total approved or in progress projects</td>
<td>20</td>
<td>26</td>
<td>40</td>
<td>64</td>
<td>70</td>
<td>75</td>
<td>107</td>
<td>106</td>
<td>82</td>
<td>85</td>
<td>20</td>
</tr>
</tbody>
</table>

Table from Michael Challands, Stats NZ
2. What have we learned?

ii. Publications increasing year on year

https://vhin.co.nz/research/
2. What have we learned?

ii. Publications increasing year on year

IDI Publications

- Thesis, n=31
- Journal articles, n=135
- Non-government reports, n=135
- Government reports, n=188
- TOTAL, n=489

Year: 2012 (n=6) 2013 (n=14) 2014 (n=18) 2015 (n=27) 2016 (n=33) 2017 (n=47) 2018 (n=58) 2019 (n=78) 2020 (n=74) 2021 (n=106) 2022 (n=28)

- Government
- Non-government
- Journal
- Thesis
2. What have we learned?

ii. Publications increasing year on year

- Journal articles (n=135)
  - n=41 in 2021 and n=17 by April 2022 (one a week)
  - Most articles (n=111, 82%) published in international journals
  - Median impact factor of Journals = 2.7 (4.6 in 2020/21)
    - n=23 (17%) publications in Journals with impact factors>5
    - n=7 (5%) publications in Journals with impact factors>10
2. What have we learned?

iii. Few reports taking a ‘social investment approach’

- Only 10/107 reports 2015-17 have a social investment focus
  - Characteristics of Children at Greater Risk of Poor Outcomes as Adults. *The Treasury (2016)*
  - Using Integrated Administrative Data to Understand Children at Risk of Poor Outcomes as Young Adults. *The Treasury (2015)*
  - Using Integrated Administrative Data to Identify Youth Who Are at Risk of Poor Outcomes as Adults. *The Treasury (2015)*
  - Economic Outcomes of Youth Not in Education, Employment or Training (NEET). *The Treasury (2016)*

- Caveat: I may have missed/misclassified some; results of some may not have been published
iv. Publications cover diverse areas
- Labour market (Motu, MBIE, AUT, Victoria)
- Productivity (Motu, MBIE, Productivity Commission)
- Education (MOE, Motu, NZ Initiative)
- Health (Universities)
- Housing (Otago, MBIE, SWA)
- Wellbeing (Motu, Treasury, SWA, Universities)
- Deprivation (Auckland)
- Covid (Te Punaha Matatini)
- Methods (Stats NZ, Others)
Diversity of IDI investigations: COVID

- ‘Aotearoa Coincidence Network’
- Transmission risk estimated for NZ population using data on ‘connections’ between individuals (employed at the same workplaces or attend the same schools)
- Health vulnerability estimated for each SA2 by number elderly or with long term conditions

Turnbull et al. (2022). Investigating the transmission risk of infectious disease outbreaks through the Aotearoa Co-incidence Network (ACN): a population-based study. The Lancet Regional Health-Western Pacific. 20:100351
Diversity of IDI investigations: Vegetation diversity and asthma

Born 1998 (Birth records)  
→  
Area of residence 1998-2016  
(Address table)  
→  
Area-level greenness index  
(NDVI: Normalized difference vegetation index) 1998-2016  
(satellite imagery)  
→  
Asthma 2005-2016  
(Hospitalization and pharmaceutical data)

Most children born very preterm (23-24w) able to be resuscitated and 66% survived to age 10

Most able to participate in and perform well at school

Gradient of school achievement across gestational age, right up to early term (37-38w)
2. What have we learned?

vi. Comparing IDI results to ‘collected’ data can be illuminating

2. What have we learned?

**Strengths**

- Whole population
  - Transmission risk, Turnbull et al., 2022
- Able to study small and difficult to reach populations
  - pre-term babies, Berry et al., 2018; refugee children, Charania et al., 2018; 2020
- All life stages able to be studied
- Cross-domain exposures and outcomes
  - Berry et al., 2018
  - Donovan et al., 2018

**Weaknesses**

- Primarily service use data
  - Will not include problems/issues for which service has not been sought
  - May be biases in who receives services
  - Interpret with care, Svardal et al., 2021
- Primarily deficit focussed
- Limited follow-up for many datasets
- Variable data quality and documentation
- Difficult to determine family and household structures longitudinally
- Missed and incorrect links
3. What are the issues?

i. Social Licence

“permission to make decisions about management and use of the public’s data”

- Kalkman et al. 2019; Paprica et al. 2019: Public acceptance for secondary data use when there is transparency about how the data are being used, and there is trust in the institutions
- Do people know how their data are being used?
  - Unlikely… Nielsen survey found that 40% of adults didn’t know enough about what StatsNZ does (generally) to know whether or not to trust them
3. What are the issues?

ii. Ethics

- Ethical governance of IDI projects is ‘uneven’, and depends on
  - The researcher’s understanding of their obligations and the researcher’s employer/sector
    - University researchers:
      No-one forcing them, but ethical governance systems in place
    - Government researchers:
      Might have systems in place and might be required to use them
    - Other researchers:
      ?? May not have ethical governance systems in place ??
3. What are the issues?

ii. Ethics

• Informed consent often lacking
  • Individuals may have signed a form indicating their data may be used for ‘research or scientific purposes’, but…
    • Often under duress (need to use a service)
    • Unclear what public understand by ‘research or scientific purposes’
  • Lessof (2009): the important question regarding use of administrative data without informed consent is “whether an individual’s health, interests or confidentiality could be affected negatively”
    • Use of IDI data unlikely to affect health, and confidentiality is guarded (‘Five Safes’) … but… ethical assessment by diverse panel would help assess whether individuals’ interests could be affected negatively
3. What are the issues?

iii. Māori data sovereignty

“Māori data should be subject to Māori governance”

- ‘Big data divide’ – those represented in the data often not those using and benefitting from data use

- Data structured to respond to government priorities, NOT to benefit Indigenous nations (Rowe et al., 2021)
  - Ongoing community-level consent crucial

- Ngā Tikanga Paihere framework goes some way to address sovereignty concerns
iii. Māori data sovereignty

“Māori data should be subject to Māori governance”

• Other frameworks go further
  • For datasets where indigenous populations are over-represented, Manitoba Multigenerational Cohort requires approval from the First Nations Health and Social Secretariat of Manitoba and the Manitoba Metis Community Research and Ethics Protocol, in addition to an ethical assessment (Hamad et al. 2021)
  • Is the current governance arrangement of the IDI the best way to recognize Māori data sovereignty??
3. What are the issues?

iv. Inequality of access

- High bar to access IDI (Five Safes), so many communities who may benefit from IDI data not able to access it (‘big data divide’…)
  - Covid lockdowns highlighted inequality (unable to access ‘Safe Setting’)
  - Most uses of the IDI are simple ones (counts: “how many people are ‘X’?”, Ellis, 2017).
- So… is a technical solution possible that allows broader access to data, relaxing ‘Safe People’ & ‘Safe Setting’ (e.g., an extension of NZ.Stat)?
- Paprica et al., 2019: The more the public knows about the data the greater the trust and support (social licence!) for the data-use project
  - What better way to learn about the data than to use it and see its benefits! In so doing…
  - Communities’ views and aspirations better reflected in the research about them
Final thoughts

- As an academic researcher, the IDI is a fantastic research tool
  - The IDI allows research that is unable to be done any other way
- As a New Zealander, I want the public to know more about the IDI
  - to discuss its benefits and risks
  - to have a say in the research that is being undertaken
  - and to (perhaps…) participate in the research
QUESTIONS?
References (1/2)


Crichton S, Templeton R, Tumen S. 2015. Using Integrated Administrative Data to Understand Children at Risk of Poor Outcomes as Young Adults. Wellington, New Zealand: The Treasury


