Data that leaves no one behind

• What is the IDI?
• How is the IDI being used?
• Lessons learned and next steps
What is the IDI?

Integrated Data Infrastructure (IDI)

An integrated database containing de-identified longitudinal and geographical microdata about people and households

For a full list of data, go to www.stats.govt.nz/idi
A picture of people’s lives

Birth → Pre-school → School age → Adulthood

Flow of data in the IDI

1. Data collected from all sources
2. Process and link the data
3. De-identified data available for research
Linking datasets together

1. Spine created by linking births, tax, and visa data together
2. Other sources (aka nodes) are linked with the spine.

Linking records

Birth data

Tax data

Record 1

Record 52

LINK
How is the IDI being used?

The social investment approach

Getting the right help to people who need it, at the time it will make the most difference
Childhood risk factors

How can we better target services to vulnerable children?

https://shinyapps.stats.govt.nz/sii/

Childhood risk factors (2)
Investment Approach to Justice

How can crime and harm from crime be prevented?

4-16 year old serious offenders are a high risk group

They’ve had a difficult start in life

- 62% have a parent who’s had corrections management
  - HALF ARE IN A SOLE PARENT FAMILY
  - ON AVERAGE, THERE ARE 2 OTHER CHILDREN LIVING IN THE SAME HOUSEHOLD

60% are supported by a benefit
- half of the child of beneficiary compared to only 1% of all 14-16 year olds

They have broader problems

- 82% of the 14-16 year group have been notified to CYF for a care and protection matter
- connected to children who were in care as of April 1, 2016

- 45% have accessed secondary mental health services or pharmaceuticals relating to a mental health condition in the past year
- connected to 14 to 16 year olds

58% have been stood down or suspended twice or more from school
- compared to only 1% of all 14-16 year olds

What can we do?

We’re exploring options to expand investment in these areas

- COGNITIVE-BEHAVIOURAL THERAPY
- FAMILY THERAPY
- PROFESSIONAL MENTORING SERVICES

... because the evidence shows they reduce crime
Populations of interest - Maori

What can Statistics NZ do to better partner with Maori?

Hard to reach populations – vulnerable and transient people

How do we improve outcomes for the homeless?
Lessons learned – enabling factors

- Statistics Act 1975
- Stats NZ is a ‘safe pair of hands’, has had experience in data integration
- Better public services measures
- Social Investment Approach
- Ministerial champions
- Social licence
Lessons learned – implementation

• “If you build it, they will come”
• Data providers generally willing to make their data available
• Technical skills are a barrier to extracting value
• Early issues with system access and capacity
• Informed trust

A register-based statistical system
IDI Spine

- Spine:
  - a primary dataset that other datasets are linked to
  - complete list of uniquely identified members of the target population
  - maximum coverage from a minimum number of sources
  - minimum duplicates
  - high quality (low false positive rate and low false negative rate)

Prototype spine

- The target population for the prototype spine is broadly an ‘ever-resident’ population
- It includes
  - people born in NZ
  - permanent residents
  - those with visas that allow them to reside, work or study in NZ (including international students and temporary workers)
  - those living and working here without a requirement for a formal visa (e.g. Australians living in NZ)
- It excludes short term visitors (such as tourists)
Data in the spine

- Births (Department of Internal Affairs)
- Visa data (Ministry of Business, Innovation and Employment) – excluding visitor and transit visas
- Tax data (Inland Revenue)

IDI spine

- Births
  - 44%
- Visa
  - 22%
  - 8%
- IR
  - 17%
  - 0.2%
  - 0.02%
- Total spine ~ 9 million

~ 9 million
Deterministic linking

- Exact match
- Common unique identifiers
- Where ever possible we will match deterministically first

Probabilistic linking

- Consider two records and compare the values of the variables in common
- Given the quality of the variables and commonality of the values, assign a probability of these two records belonging to the same person
- The record pair with the highest probability wins
Blocking and linking variables

• Blocking variables have to be the same for pairs to be created
  - Any records where the blocking variable is in error or missing will not be considered for linking
• Linking variables allow comparison of records within blocks
  - Records are evaluated on the closeness (comparison function) of their linking variables
• The weights (determined from the quality and the commonality for the linking variables) form a distribution

Typical set of passes

<table>
<thead>
<tr>
<th>Pass</th>
<th>Blocking variables</th>
<th>Linking variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Date of birth</td>
<td>First names, Last names, and Sex</td>
</tr>
<tr>
<td>2</td>
<td>Soundex first names</td>
<td>First names, Last names, Sex, Year of birth, Month of birth, Day of birth</td>
</tr>
<tr>
<td></td>
<td>Soundex last names</td>
<td></td>
</tr>
</tbody>
</table>

• Soundex is a phonetic encoding of a string that helps account for spelling variables
  eg Katrina and Katherine are coded K365
False positives and negatives

- True matches
- Links made
- False negatives
- False positives

False positives

- Link made is not a true match
- Assessed by clerical review
- Sample of a few hundred
- Normally keep false positive rate under 2%
- Aim for <1% false positive rate for the spine
Common reasons for false positives

• Linking methodology means increased weights don’t always mean better quality
  • How short names are dealt with in comparison functions eg. Si v Smith
  • A rare first name with differing surnames may get a very high weight but we may call it a false positive

What causes false negatives?

• Changed or missing names
• Choices about linking strategy
• Limitations in our linking methods

<table>
<thead>
<tr>
<th>Dataset 1</th>
<th>Dataset 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elizabeth Bennet</td>
<td>Elizabeth Darcy</td>
</tr>
<tr>
<td>James Alan Rodney Smith</td>
<td>Jim Smith</td>
</tr>
<tr>
<td>DOB 03/10/1975</td>
<td>DOB 10/03/1975</td>
</tr>
<tr>
<td>&lt;BLANK&gt; Johnson, Age 50</td>
<td>Patrick Phillip Johnson, 15/9/1965</td>
</tr>
</tbody>
</table>
We operate within a ‘five safes’ framework to ensure that access to the IDI is only provided if all of the following conditions can be met:

1. Safe people
2. Safe projects
3. Safe settings
4. Safe data
5. Safe output
Access to the IDI for research

**Go**
- You need five green before microdata access can be considered
  - Research is for a statistical purpose
  - Research is for the public good
  - Research will be conducted by a credible team
  - Suitable data is available
  - Statistics NZ can enforce an agreement

**Check**
- You need to talk to us to check each orange issue for potential problems
  - Research is for commercial gain
  - Unrecognised research institution
  - Data integration is required
  - Researchers are based overseas
  - Quality of data is not adequate for research purpose
  - Lead researcher is a student or junior staff member
  - Research is for service provider evaluation
  - Non-government researchers are requesting access to business tax data

**Stop**
- You need to talk to us to resolve all red issues before microdata access can be considered
  - Research is about named people or businesses
  - Research is for a regulatory purpose
  - Researchers do not have organisational support
  - Research will not be released
  - Complex legislative issues affect access

---

**Before applying**
- Website: stats.govt.nz/idi
- Talk to us
- Researchers must be: Capable, Trustworthy and in NZ
- Research must have: strong public good element

**Your application**
- Talk to us
- Health projects need to check if ethics approval needed
- Some data sets accesses need agency review

**In the data lab**
- You’ll sign a researcher undertaking, declaration of secrecy and contract
- New researchers start at Statistics NZ Data Lab (confidentiality training)
- Once approved your IDI user profile is built and accesses allocated
- Output is checked to ensure correct confidentiality rules have been applied

**Support**
- The wiki within the secure Data Lab environment
- Online researcher and agency forum: MeetaData
- Ongoing forums. You are encouraged to be part of the community
What does the NZ public think?

Social licence

“Social licence is societal acceptance that a practice that lies outside general norms may be performed by a certain agent, on certain terms. It is the result of a process of negotiation with a wider societal group, and means that the practice can be performed by that agent without incurring social sanction.”

From Surveys, social licence and the IDI, a collaborative project between the University of Auckland and Statistics New Zealand, co-funded by the Data Futures Partnership.
“Aren’t you doing that already?”

- Public attitudes to data integration report online
- Moving from naïve trust to informed trust

<table>
<thead>
<tr>
<th>Less acceptable</th>
<th>More acceptable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ambiguous and general</td>
<td>Purpose</td>
</tr>
<tr>
<td>Insecure</td>
<td>Transparent and specific need</td>
</tr>
<tr>
<td>Insecure</td>
<td>Security</td>
</tr>
<tr>
<td>Insecure</td>
<td>Secure</td>
</tr>
<tr>
<td>Open access</td>
<td>Access</td>
</tr>
<tr>
<td>Open access</td>
<td>Strict restrictions, procedures, and</td>
</tr>
<tr>
<td></td>
<td>protocols</td>
</tr>
<tr>
<td>Public/personalised</td>
<td>Privacy</td>
</tr>
<tr>
<td>Public/personalised</td>
<td>Confidential/anonymous</td>
</tr>
<tr>
<td>Private</td>
<td>Interests</td>
</tr>
<tr>
<td>Private</td>
<td>Public</td>
</tr>
<tr>
<td>Subjective and biased</td>
<td>Data and metrics</td>
</tr>
<tr>
<td>Subjective and biased</td>
<td>Representative and meaningful</td>
</tr>
<tr>
<td>Unfair and harmful</td>
<td>Outcomes</td>
</tr>
<tr>
<td>Unfair and harmful</td>
<td>Fair and beneficial</td>
</tr>
</tbody>
</table>

Social licence

- Building public acceptability and informed trust

<table>
<thead>
<tr>
<th>Less acceptable</th>
<th>More acceptable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ambiguous and general</td>
<td>Purpose</td>
</tr>
<tr>
<td>Insecure</td>
<td>Transparent and specific need</td>
</tr>
<tr>
<td>Insecure</td>
<td>Security</td>
</tr>
<tr>
<td>Insecure</td>
<td>Secure</td>
</tr>
<tr>
<td>Open access</td>
<td>Access</td>
</tr>
<tr>
<td>Open access</td>
<td>Strict restrictions, procedures, and</td>
</tr>
<tr>
<td></td>
<td>protocols</td>
</tr>
<tr>
<td>Public/personalised</td>
<td>Privacy</td>
</tr>
<tr>
<td>Public/personalised</td>
<td>Confidential/anonymous</td>
</tr>
<tr>
<td>Private</td>
<td>Interests</td>
</tr>
<tr>
<td>Private</td>
<td>Public</td>
</tr>
<tr>
<td>Subjective and biased</td>
<td>Data and metrics</td>
</tr>
<tr>
<td>Subjective and biased</td>
<td>Representative and meaningful</td>
</tr>
<tr>
<td>Unfair and harmful</td>
<td>Outcomes</td>
</tr>
<tr>
<td>Unfair and harmful</td>
<td>Fair and beneficial</td>
</tr>
</tbody>
</table>
Benefits of integrated data

- We can tackle previously ‘unanswerable’ questions
- Longitudinal view
- Cross sector view
- Geographical views
- Reduced research cost and burden

Limitations of the IDI

- Capability
- Finding a study population
- Establishing comparable population
- Administrative data quality issues
- Small number studies limitations
- High time and tech investment (currently)
Where to next

- **Analysis**
  - Lifecourse analysis, actuarial modelling, programme evaluation, micro-simulation modelling

- **Infrastructure**
  - Cloud-based platform, sustainability

- **Boost use**
  - Provide products and services, researcher toolbox, improve access pathways

- **Supporting system**
  - Improve data quality, grow data analytic capability, improve sharing infrastructure and standards, testing legislation, ethics approvals
Where to next - Analysis

- Types of analyses users want to do are becoming more complex, e.g. lifecourse actuarial modelling
- Issues from a capability perspective but also a structural basis
- Quick turn-around policy analysis – simple analysis on complex data

Where to next - Infrastructure

- Cloud-based platform
- Sustainability
Where to next – Boost use

- Provide products and services
- Researcher toolbox
- Improve access pathways

Where to next – Supporting system

- Improve data quality
- Grow data analytic capability
- Improve sharing infrastructure and standards
- Testing legislation
- Ethics of big data
Unleashing the power of data to change lives

History of the IDI

- 1997 Stats NZ began performing data integration
- 2011 IDI prototype established
- 2012 Government launched Better Public Services
- 2013 Expansion of the IDI
- Further growth continues…
Virtual Health Information Network

What is the cost of cardiovascular disease on our health system?

Temporary migration

Should international students be allowed longer work entitlements?
Health Pathways and Outcomes

What is the impact of mental health on education, employment and training?

NEETs

What happens to school leavers who don’t go on to further education or employment?