The Integrated Data Infrastructure (IDI):
New Zealand's Bold Data Experiment

**Webinar, ISPOR
6 July 2017**

COMPASS
RESEARCH CENTRE

FACULTY OF ARTS
THE UNIVERSITY OF AUCKLAND

Whare Wānanga o Tāmaki Makaurau

Barry Milne

COMPASS Research Centre

University of Auckland

Outline

- **Integrated Data Infrastructure (IDI)**
  - What is it?
  - Where did it come from?

- **Use of the IDI**
  - Who is using it and how?
  - Virtual Health Information Network

- **Privacy and risk mitigation**
  - Privacy protocols
  - Managing threats to privacy

Integrated Data Infrastructure (IDI)

Integrated Data Infrastructure

The Integrated Data Infrastructure (IDI) is a large research database containing microdata about people and households. Data is from a range of government agencies, Statistics NZ surveys including the 2013 Census, and non-government organisations. The IDI holds over 166 billion facts, taking up 1.22 terabytes of space – and is continually growing. Researchers use the IDI to answer complex questions to improve outcomes for New Zealanders.
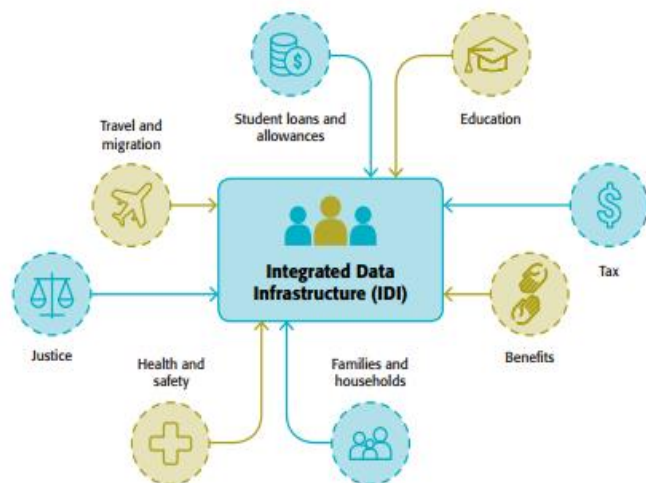
About the Integrated Data Infrastructure

The Integrated Data Infrastructure (IDI) combines information from a range of organisations (such as health and education data) to provide the insights government needs to improve social and economic outcomes for New Zealanders.

With all personal information removed, integrated data gives a safe view across government so agencies can deliver better services to the public and ensure investment is made where it's needed most. Integrated data is particularly useful to help address complex social issues such as crime and vulnerable children.

New Zealand

The University of Auckland

Data in the IDI October 2016

Statistics
New Zealand
TATAURANGA AOTEAROA

Statistics New Zealand's Integrated Data Infrastructure (IDI) is a large research database containing de-identified microdata about people and households.

Integrated Data Infrastructure (IDI)

Travel and migration · Student loans and allowances · Education · Justice · Health and safety · Families and households · Benefits · Tax

The IDI contains person-centred microdata from a range of government agencies, Statistics NZ surveys including the 2013 Census, and non-government organisations. For more information about data in the IDI, see **www.stats.govt.nz/idi-data**.

The Longitudinal Business Database (LBD) complements the IDI with microdata about businesses. For more information about data in the LBD, see **www.stats.govt.nz/lbd**.

Health and safety data

- ACC injury claims – from 1994
- B4 School Checks – from 2011
- Cancer registrations – from 1995
- Chronic conditions – from 2007
- General medical services claims – from 2002
- Health tracker – 2006–13
- Laboratory claims – from 2003
- Mortality – from 1988
- Immunisation – from 2006
- National non-admitted patient collection – from 2007
- Pharmaceuticals – from 2005
- PHO enrolments – from 2003
- Population cohort demographics and addresses – from 2004
- Mental health and addiction – from 2008
- Publicly funded hospital discharges – from 1988
- National Needs Assessment and Service Coordination Information System (SOCRATES)

Justice data

- Recorded crime: offenders – from 2009
- Recorded crime: victims – from 2014
- Court charges – 1992–2013
- Sentencing and remand – from 1998

Benefits and social services data

- Benefits – from 1990
- Youth services – from 2004
- Auckland City Mission – from 1996
- Children's Action Plan – from 2013

Tax and income data

- Tax and income – from 1999

Education and training data

- Early childhood education – 2008–15
- Primary education – from 2007
- Secondary education – from 2004
- Tertiary education – from 1994
- Industry training – from 2001
- Targeted training – from 2001

Student loans and allowances data

- Student loans and allowances – from 1992

Travel and migration data

- Driver licence and motor vehicle registers
- Border movements – from 1997
- Visa applications – from 1997
- Departure and arrival cards – from 1997
- Migrant Survey – from 2012
- Longitudinal Immigration Survey of NZ – 2005–09

Family and household data

- 2013 Census
- Births, deaths, marriages, and civil unions – from 1840
- Child, Youth and Family – from 1991
- Household Economic Survey – from 2006
- Household Labour Force Survey – from 2006
- NZ Income Survey – from 2006
- Working for Families – from 2003
- Tenancy – from 2000
- Social housing – from 1980
- Survey of Family Income and Employment – 2002–10
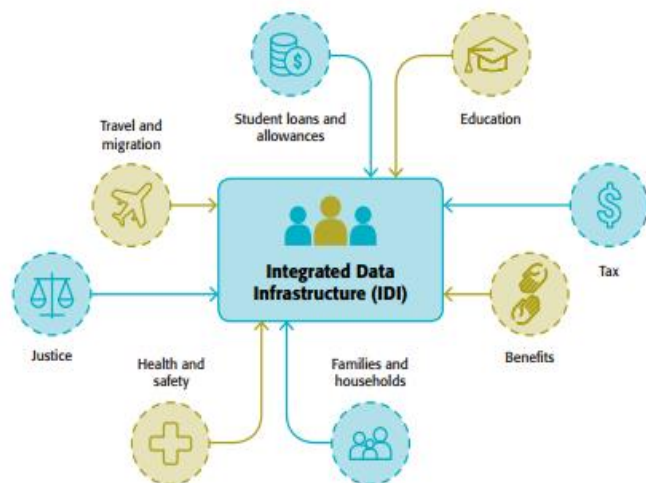
History of NZ Integrated Data

- ■ SNZ Integration projects prior to IDI
  - Student Loans and Allowances integrated Dataset
    - IRD, MSD (StudyLink), MoE
  - Linked Employer-Employee Data (LEED)
    - IRD, SNZ Business Data Frame
  - LEED-MSD Benefit Dynamics Data
  - Employment Outcomes of Tertiary Education Study
    - LEED-MoE
  - LEED-Household Labour Force Survey

New Zealand

The University of Auckland

Use of the IDI

Data in the IDI October 2016

Statistics New Zealand
New Zealand
TATAURANGA AOTEAROA

Statistics New Zealand's Integrated Data Infrastructure (IDI) is a large research database containing de-identified microdata about people and households.

The IDI contains person-centred microdata from a range of government agencies, Statistics NZ surveys including the 2013 Census, and non-government organisations. For more information about data in the IDI, see **www.stats.govt.nz/idi-data**.

The Longitudinal Business Database (LBD) complements the IDI with microdata about businesses. For more information about data in the LBD, see **www.stats.govt.nz/lbd**.

Health and safety data

- ACC injury claims – from 1994
- B4 School Checks – from 2011
- Cancer registrations – from 1995
- Chronic conditions – from 2007
- General medical services claims – from 2002
- Health tracker – 2006–13
- Laboratory claims – from 2003
- Mortality – from 1988
- Immunisation – from 2006
- National non-admitted patient collection – from 2007
- Pharmaceuticals – from 2005
- PHO enrolments – from 2003
- Population cohort demographics and addresses – from 2004
- Mental health and addiction – from 2008
- Publicly funded hospital discharges – from 1988
- National Needs Assessment and Service Coordination Information System (SOCRATES)

Justice data

- Recorded crime: offenders – from 2009
- Recorded crime: victims – from 2014
- Court charges – 1992–2013
- Sentencing and remand – from 1998

Benefits and social services data

- Benefits – from 1990
- Youth services – from 2004
- Auckland City Mission – from 1996
- Children's Action Plan – from 2013

Tax and income data

- Tax and income – from 1999

Education and training data

- Early childhood education – 2008–15
- Primary education – from 2007
- Secondary education – from 2004
- Tertiary education – from 1994
- Industry training – from 2001
- Targeted training – from 2001

Student loans and allowances data

- Student loans and allowances – from 1992

Travel and migration data

- Driver licence and motor vehicle registers
- Border movements – from 1997
- Visa applications – from 1997
- Departure and arrival cards – from 1997
- Migrant Survey – from 2012
- Longitudinal Immigration Survey of NZ – 2005–09

Family and household data

- 2013 Census
- Births, deaths, marriages, and civil unions – from 1840
- Child, Youth and Family – from 1991
- Household Economic Survey – from 2006
- Household Labour Force Survey – from 2006
- NZ Income Survey – from 2006
- Working for Families – from 2003
- Tenancy – from 2000
- Social housing – from 1980
- Survey of Family Income and Employment – 2002–10

Statistics New Zealand operates a five-safes environment, balancing privacy and confidentiality with data insights.

For information about applying to use the IDI or to learn about how we keep the data safe, see **www.stats.govt.nz/idi**

Who is using the IDI?

- 117 IDI projects listed on StatsNZ website

- 64 led (co-led) by government departments
  - MBIE: 13; MSD: 10; MOE: 8

- 41 led (co-led) by Universities
  - UOA: 13; Otago: 12 ; AUT: 7

- 12 led (co-led) by other groups
  - Motu:5

http://www.stats.govt.nz/browse_for_stats/snapshots-of-nz/integrated-data-infrastructure/researchers-using-idi.aspx

Who is using the IDI?

	GOVT	UNIV	OTHER	TOTAL
Benefits and Social Services	10	1	1	**12**
Business and Employment	15	6	3	**24**
Education and Training	11	2	2	**15**
Family and Households	5	12	1	**18**
Health and Safety	8	18	2	**28**
Housing	4	1	2	**7**
Justice	4	0	0	**4**
Travel and Migration	7	1	1	**9**
TOTAL	**64**	**41**	**12**	**117**

New Zealand

The University of Auckland

9

Health projects in the IDI

- ▶ **MAA2016-44 Long term outcomes for peri-viable births**

- ▶ **MAA2016-21 The health impacts of playing Rugby**

- ▶ **MAA2016-10 Environmental and occupational risk factors for chronic conditions**

- ▶ **MAA2015-42 Impact of the MeNZB™ vaccine on gonorrhoea**

- ▶ **MAA2015-41 Modelling the cost effectiveness of changes to income tax on obesity prevention perspective in New Zealand**

- ▶ **MAA2014-22 The impacts of cancer, chronic disease, and acute health events on future employment, earnings, and benefit receipt among the working-age population**

- ▶ **MAA2016-12 Using birth information to predict reaching key early childhood development indicators: identifying at risk populations**

Outcomes of tertiary study

Impact of health conditions

COMPASS
RESEARCH CENTRE

FACULTY OF ARTS
THE UNIVERSITY OF AUCKLAND

Whare Wānanga o Tāmaki Makaurau

➡ What impacts do eight selected health conditions have on the **employment rates** of working adults who experience them?

Research undertaken by Silvia Dixon, Treasury

12
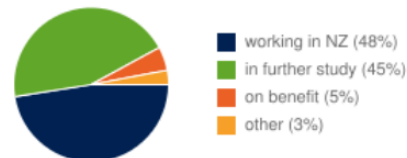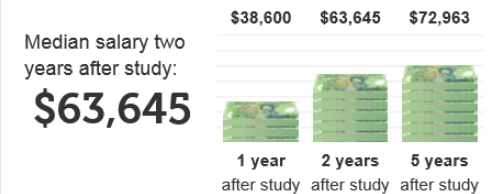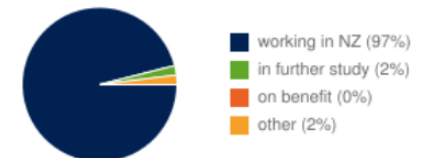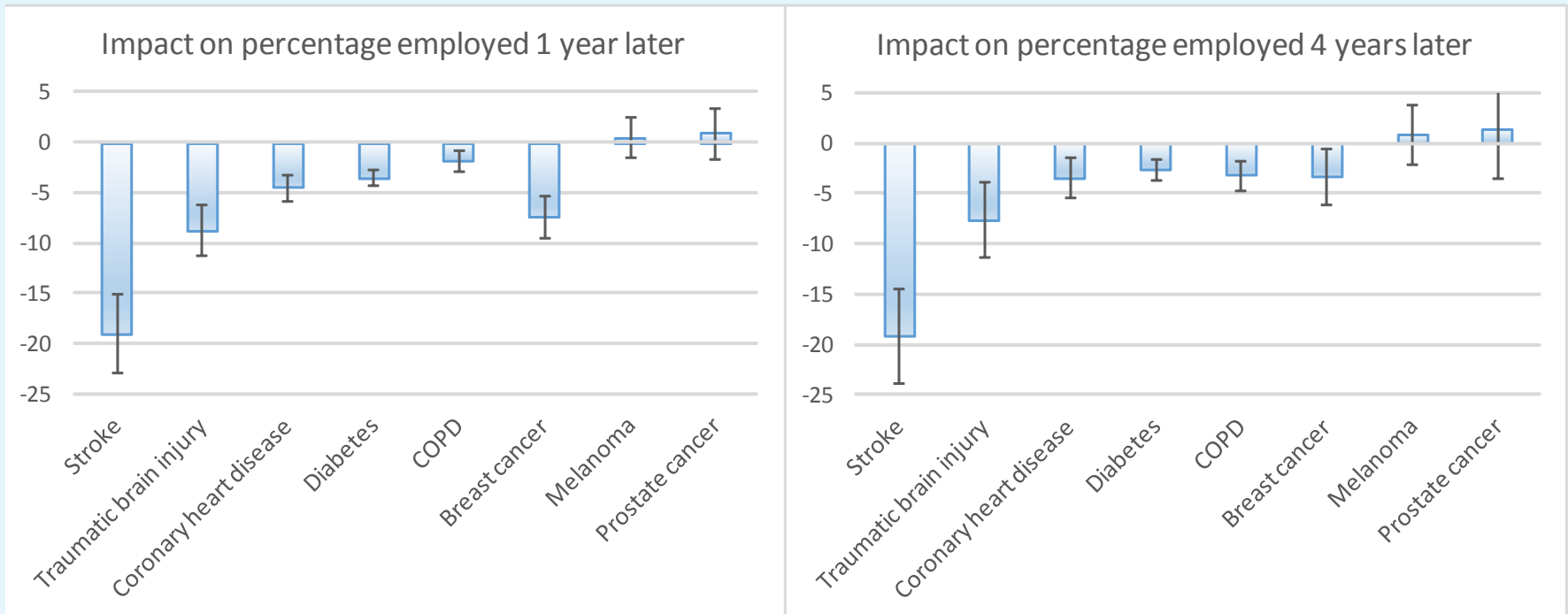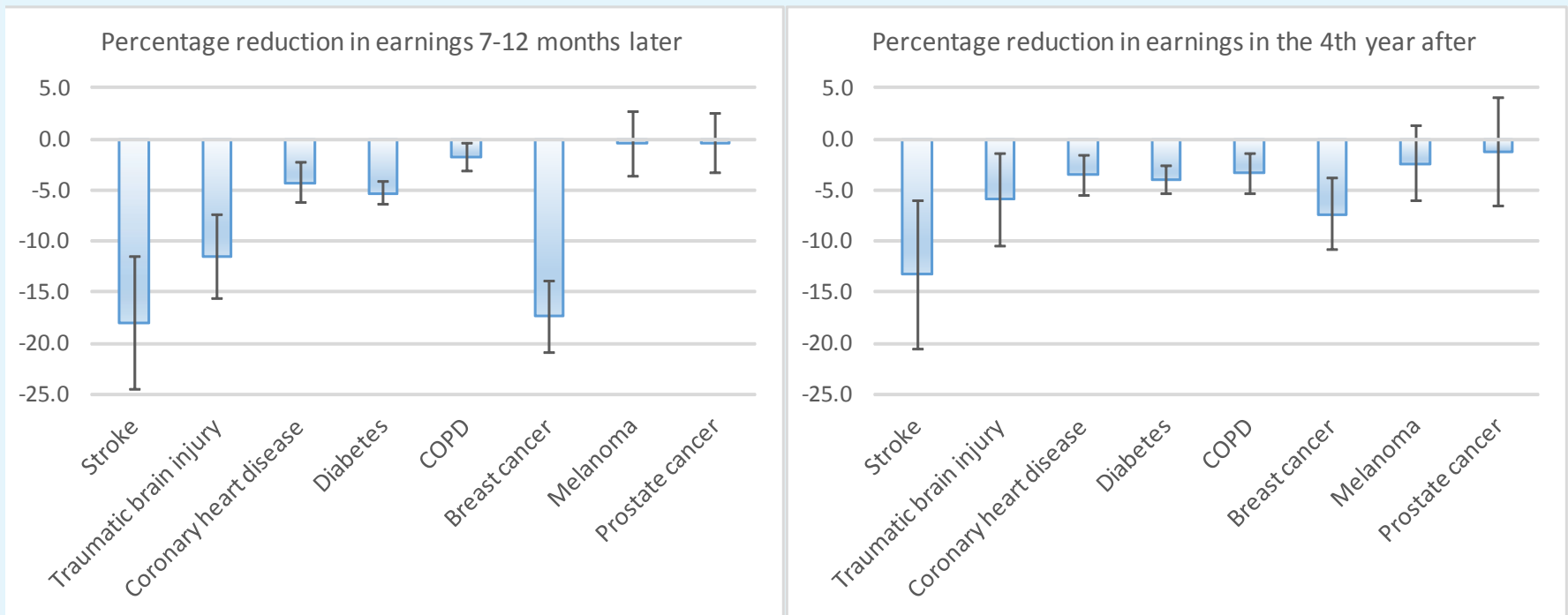
Impact of health conditions

➡ What impacts do eight selected health conditions have on the **earnings** of working adults who experience them?

Percentage reduction in earnings 7-12 months later

Percentage reduction in earnings in the 4th year after

Conditions (both charts): Stroke, Traumatic brain injury, Coronary heart disease, Diabetes, COPD, Breast cancer, Melanoma, Prostate cancer

New Zealand

The University of Auckland

Research undertaken by Silvia Dixon, Treasury

13

Better Start National Science Challenge

- The Better Start Challenge focuses on three critical areas of childhood development that have been linked to life course outcomes.

- These are:
  - Obesity
  - Literacy
  - Mental Health
  - Autism Spectrum Disorder (ASD)

Better Start National Science Challenge

Investigations

- Obesity, ethnicity and deprivation over time
- Unpacking the relationship between obesity, deprivation and ethnicity
- Trajectories in obesity across Territorial authorities
- Community Resources and obesity
- Who's not doing the B4SC
- Obesity and Lift the Lip oral health assessment
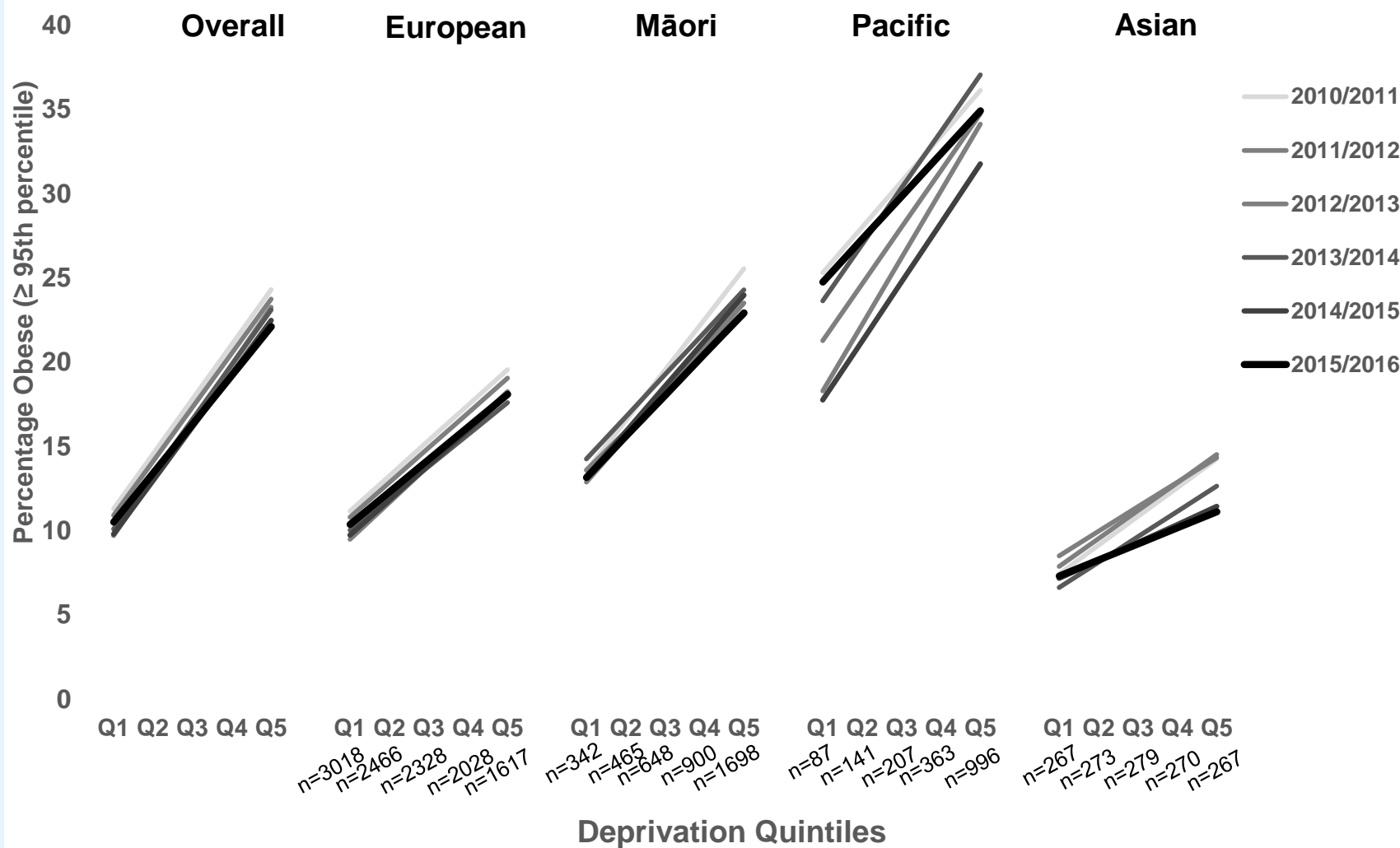- Seasonal trends in obesity
- SSRI use and ASD Prevalence

New Zealand

The University of Auckland

Obesity, ethnicity and deprivation over time

Risk factors for congenital malformations

Andrea t' Mannetje (Massey)

- ~2500 infants diagnosed per year, 20% of all infant deaths in NZ

- Modifiable risk factors have not been studied previously in NZ

- Overseas studies have suggested that pharmaceutical, occupational exposures may be risk factors, but sample sizes are small

New Zealand

The University of Auckland

Risk factors for congenital malformations

- 3000 babies with CM born in 2007-2009 selected from NZ Birth Defects Monitoring Programme

- Control babies selected from maternity data

- 600 case and 600 control mothers interviewed

- Can information be obtained about the others by linking with IDI?

New Zealand

The University of Auckland

Risk factors for congenital malformations

New Zealand

The University of Auckland

```
                          Total
                          5745
                            |
            ┌───────────────┴───────────────┐
   Linked to birth record          Not linked to birth record
          5664                               81
          (99%)                              (1%)
            |
   ┌────────┴────────┐
Mother on birth record    Father on birth record
      5664                       5337
      (99%)                      (93%)
      |
Mother has census record
      4320
      (75%)
```

Case: 77%
Control: 75%

19
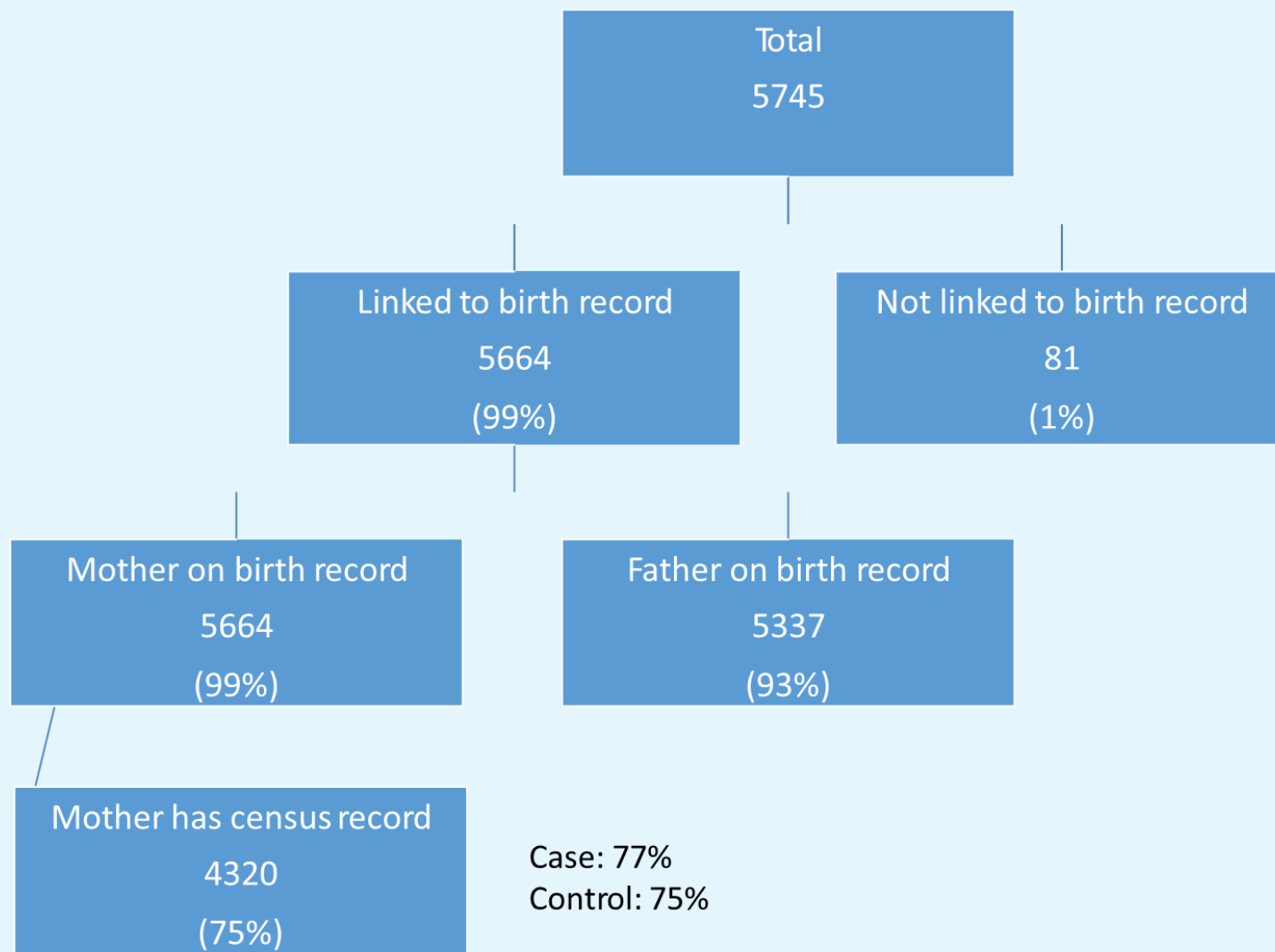
Folate antagonists

	OR (95% CI)		
	All CM	Circulatory	Musculoskeletal
3 months preconception	1.9 (1.2, 2.9)	2.8 (1.6, 4.9)	2.4 (1.3, 4.5)
First trimester	2.2 (1.3, 3.7)	2.7 (1.4, 5.4)	2.9 (1.4, 6.0)
Second trimester	2.1 (1.1, 4.0)	2.6 (1.1, 5.9)	2.6 (1.1, 6.3)
Third trimester	1.3 (0.8, 2.1)	1.0 (0.5, 2.2)	1.1 (0.5, 2.5)

Adjusted for baby's sex, mother's age, ethnicity, quals,
smoking, nzdep, father on birth certificate
Most other CM types were suppressed due to small counts

Other medications

	All CM OR (95% CI)
Diabetes medications	
3 months preconception	2.5 (1.4, 4.8)
First trimester	1.9 (1.0, 3.4)
Second trimester	2.6 (1.5, 4.5)
Third trimester	1.7 (1.2, 2.4)
Epilepsy medications (any trimester)	1.6 (1.1, 2.2)

Adjusted for covariates as previous
Diabetes effect may be due to diabetes rather than medications
Small counts are still a problem

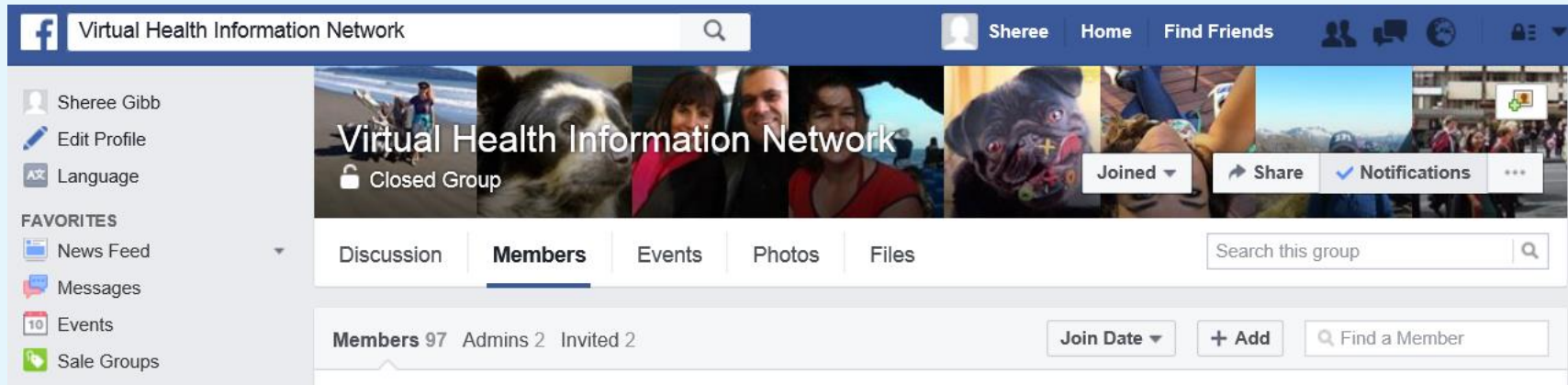# Virtual Health Information Network (www.vhin.co.nz)

The Virtual Health Information Network (VHIN) assists researchers to do high quality health research using the Integrated Data Infrastructure (IDI) and other administrative health data. A deeper understanding of the IDI and increased awareness of data quality issues can be gained via website guides (posted below), online discussions on the VHIN Facebook page, analytical code available from VHIN research and meetings about current IDI research (Presentations).

The VHIN is a network of researchers, analysts and other professionals who use health data to generate insights that support the health and wellness of all New Zealanders. The network facilitates sharing and collaboration amongst network members in order to enhance health research outputs and improve health service delivery and health outcomes in New Zealand (About VHIN).

Sponsored by

JOINT INITIATIVE

- University of Otago (Tony Blakely)

- University of Auckland (Barry Milne)

- Massey University (Jeroen Douwes)

- Ministry of Health (Kendra Telfar)

- SNZ representation

- Te Mana Rāraunga representation

- Core funding: Healthier Lives NSC

22

VHIN Facebook

- Request to join Facebook Group
- Discussions, queries & upcoming events
- @vhin_network
- vhin@otago.ac.nz

VHIN code sharing

- ➡ IDI wiki
- ➡ MeetaData
    - ➡ for access via RealMe email meetadata@stats.govt.nz
- ➡ Websites: SNZ & vhin.co.nz /

Privacy

SNZ
Data Integration Policy

Data integration policy

Policy statement

Statistics New Zealand will integrate data from separate sources when necessary to ensure we can efficiently produce the information New Zealand needs to grow and prosper.

Data integration is when two or more unit record datasets, which were originally collected for different purposes, are linked together.

We will consider integrating data to produce official statistics and related research only when all four of the following principles are met.

Principle 1: The public benefits of integration outweigh both privacy concerns about the use of data and risks to the integrity of the Official Statistics System, the original source data collections, and/or other government activities.

Principle 2: Integrated data will only be used for statistical or research purposes.

Principle 3: Data integration will be conducted in an open and transparent manner.

Principle 4: Data will not be integrated when an explicit commitment has been made to respondents that prevents such action.

All data integration activities must also comply with Statistics NZ's Security Policies Framework and our methodological standards for confidentiality, as well as the Statistics Act 1975, the Privacy Act 1993, the Public Records Act 2005, and any other legislation relevant to the source datasets.

Privacy protocols

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
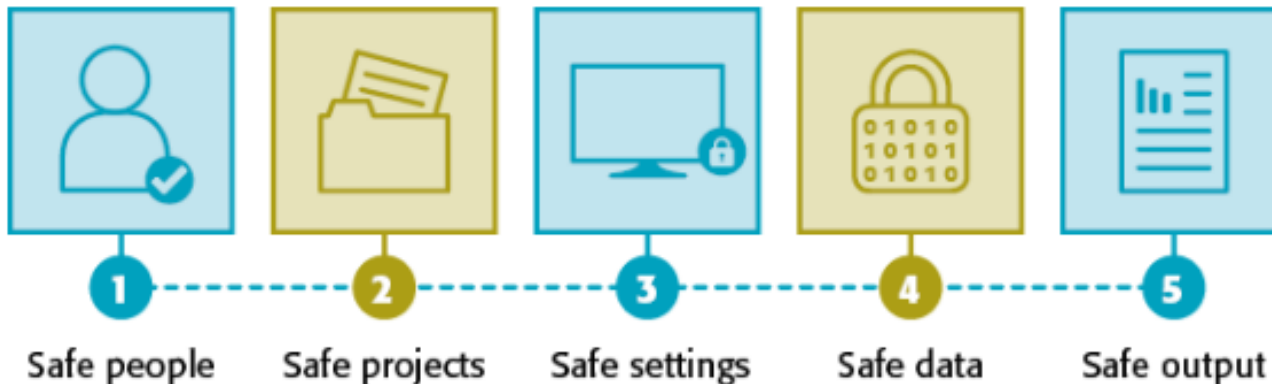
It's not about 'you' as an individual

Your personal information is never seen by researchers. This is done in two ways. First – the information is anonymised. Personal identifying information like names, addresses, and days of dates of birth are removed. Second – all research findings are confidentialised. This means that information is grouped in a way which makes it impossible to identify individual people.

New Zealand

The University of Auckland

27

Privacy protocols

- Nine privacy impact assessments undertaken (to date)

Latest privacy impact assessment

Privacy impact assessment for adding Children's Action Plan data to the IDI (published November 2016)
Benefits and risks of adding Children's Action Plan data to the Integrated Data Infrastructure.

Previous assessments

- Privacy impact assessment for adding Auckland City Mission data to the IDI (published October 2016)
  Benefits and risks of adding Auckland City Mission data in the Integrated Data Infrastructure.

- Integrated Data Infrastructure extension: Privacy impact assessment (7th ed)
  The IDI was extended in July 2016 to include transport data.

- Integrated Data Infrastructure extension: Privacy impact assessment (6th ed)
  The IDI was extended in April 2016 to include Household Economic Survey (HES) data.

- Integrated Data Infrastructure extension: Privacy impact assessment (5th ed)
  The IDI was extended in February 2016 to include social housing data and youth services data.

- Integrated Data Infrastructure extension: Privacy impact assessment (4th ed)
  The IDI was extended in October 2015 to include 2013 Census data and additional injury and offender data.

- Integrated Data Infrastructure extension: Privacy impact assessment (3rd ed)
  The IDI was extended in May 2015 to include life event data; Child, Youth and Family data; and additional health data.

- Integrated Data Infrastructure extension: Privacy impact assessment (2nd ed)
  The IDI was extended in 2014 to include selected health system data, tenancy and housing data, and additional education, Inland Revenue, and benefit data.

- Privacy impact assessment for the Integrated Data Infrastructure
  Identifies privacy risks associated with the Integrated Data Infrastructure prototype and the IDI delivered in December 2012, and outlines the processes for managing them.

Updated 4 November 2016

Public attitudes

Public attitudes to data integration

Public attitudes to data integration presents research by Opus International Consultants Ltd on behalf of Statistics NZ.

The main findings are:

- participants generally expected information provided to government departments to be shared with other government departments, and that this would be automatically and electronically conducted

- people appeared to gauge acceptability primarily on the need for the information. That is, how the data would be used and by whom

- people were interested in the value of data integration, whether the benefits would outweigh the costs and risks, and how privacy and other risks might be mitigated

- people felt that integrated data systems could be more reliable, current, and accurate than those currently used, and could result in more informed, fair, efficient, and effective decision-making and service provision.

… But social licence is an ongoing process

Identified privacy risks

Risk	Likelihood of occurring	Consequence if it did occur	Risk rating
Individuals being re-identified in the data	Unlikely	Major	Medium
Unfavourable public perception of the data integration	Possible	Moderate	Medium
Maintaining data security	Unlikely	Severe	High
Data used for non-approved purposes	Rare	Major	Low

Privacy Risk Mitigation

1. ## Re-identification

   - Removal of identifiers, 'Safe Data'

Variables that must be removed from research datasets include:

- first names

- middle names

- last names

- titles

- business names

- day of date of birth

- day of date of death

- day of date of disposal

- day of date last seen

- address information.

Privacy Risk Mitigation

1. ## Re-identification

   - Bone-fide researchers, 'Safe People'

Researchers using the IDI data will only have access to the specific datasets required for their research questions. Processes are in place to assess a potential researcher's integrity and experience. Once researchers are approved they must complete training in applying confidentiality methods and sign both a Declaration of Secrecy and a Researcher Undertaking before they can access the data.

The undertaking includes an agreement to:

- not attempt to identify particular persons or organisations

- not attempt to match the information with any other unit record level data source or list of persons or organisations

- not disclose, either directly or indirectly, information protected by the Statistics Act 1975 with any individual not approved by Statistics NZ

- apply confidentiality measures to all output to ensure that no individual person or organisation can be identified

Privacy Risk Mitigation

1. ## Re-identification

   - Confidentialised Output: 'Safe Output'

   - Counts
     - Small cell counts suppressed (<6): Can never release attribute information about a single individual, or even a small number of individuals
     - All cells 'random rounded to base 3'

   - Other
     - Income means to nearest $100 only
     - Minimum and maximums not released unless >=6 have those values

New Zealand

The University of Auckland

Privacy Risk Mitigation

1. ## Re-identification

   - ### DataLab Environment: 'Safe Settings'
     - Cannot transfer or email files outside of datalab environment (e.g., to another computer, usb drive)
     - No printing, photography
     - Non-approved people not allowed in datalab environment

   - ### Research is for statistical purposes, and is a public good research project: 'Safe projects'
     - All projects approved by Government Statistician (independent of Government by law)
     - Projects encouraged to seek ethical approval

Re-identification Disasters

- Massachusetts Group Insurance Commission released anonymised health data on state employees (1997)
  - To enable research to improve healthcare
  - Mass Governor assured public the privacy was protected by removal of identifiers
- MIT CompSci grad student, Latanya Sweeney
  - Accessed the health data
  - Accessed electoral roll ($20) for Cambridge, Mass (where Governor lived), incl name, address, ZIP code, birthdate and sex of every voter
  - Six in Cambridge shared Gov's birthdate; only three were men and only one lived in his ZIP code
  - She mailed all of the Gov's health records to him…

Re-identification Disasters

COMPASS
RESEARCH CENTRE

FACULTY OF ARTS
THE UNIVERSITY OF AUCKLAND

Whare Wānanga o Tāmaki Makaurau

As part of the investigation, John Templon, an investigative data reporter for BuzzFeed News, spent more than a year analyzing 26,000 professional men's matches and found 15 players who lost matches with unusual betting patterns "startlingly often." (Match-fixing is also believed to occur in professional women's tennis, but the BuzzFeed-BBC investigation focused only on men's tennis, so we are in this article, too.) BuzzFeed and BBC didn't name these players, citing a lack of evidence of wrongdoing and possible alternative explanations for underperformance, including injury. But BuzzFeed did release an anonymized version of the data it used on GitHub, including a file containing betting odds and the year for 129,271 matches.

Quickly, people wrote on Twitter and on GitHub that the data could be de-anonymized, thereby identifying the 15 players Templon mentioned. Ian Dorward, a London-based tennis bettor who used to set and adjust tennis betting lines for a bookmaker, emailed me the list of what he believed to be the 15 names. After Chris Bol, a data analyst based in Utrecht, the Netherlands, published the same names, Dorward went public with his findings, which criticized BuzzFeed for making the data relatively easy to crack.[1]

New Zealand

The University of Auckland

Re-identification Disasters

- Care.Data Initiative (UK, 2014)
  - Detailed NHS data made available to researchers
- Privacy campaigners noted
  - If details of an individual's treatment were in the public domain (e.g., Ed Milliband's nose operation), that individual identifiable in the dataset, and <u>all</u> treatments within NHS able to be revealed for that individual

- In the case of IDI, this would be mean all health details, tax details, benefit details, justice details, education details…
  - The dangers of n=1…
  - And the importance of SNZ's safes…

Summary

- The IDI is a large, whole population dataset made up of constituent data from govt agencies, linked at the person level

- Use to date has mainly been led by govt agencies, but use by academic researcher is increasing

- Robust systems are in place to protect the privacy of individuals and public confidence in the initiative
    - But one screw up could ruin it for everyone….

38

QUESTIONS?