Big Health Data for Epidemiology: opportunities and challenges using the SNZ IDI and other sources of "big data" for epidemiological research in NZ

George Disney (convener today) and Kate Sloane (administrator)

Tony Blakely
Director: BODE³, NZ Census-Mortality Study and CancerTrends
Co-Champion (with Osborne, Douwes and Milne; Ross): Virtual Health Information Network
Lead investigator: Healthier Lives NSC-funded VHIN
Presentations

- Introduction to day
- Introduction to VHIN
- CVD health system cost catalyst project
- Methodological opportunities
Introduction

• Welcome from Bridget Kool
• Brief overview of institutions represented today
• Objectives of the day
• Overview of timetable
Objectives of the day

• Learn about and discuss emerging opportunities for research using routine data, privacy issues, Māori data governance, etc

• Hear about current research (and the research agendas), collaborations, research questions, methods, collaborative tools for sharing code, meta data, and derived variables for research:
  – Conducted in the SNZ Integrated Data Infrastructure (IDI)
  – Conducted outside the IDI (e.g. PREDICT/VIEWS, Pharmacoepidemiology Research Network, Injury Prevention Research Unit)
  – Many more examples – apologies to those not included

• Learn about, and contribute to the evolution of, the Virtual Health Information Network (VHIN)
<table>
<thead>
<tr>
<th>Session (Chair)</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>0915 – 1030 (Barry Milne)</strong></td>
<td><strong>Integrated Data Infrastructure:</strong> SNZ and the Ministry of Health – Anna McDowell and Simon Ross</td>
</tr>
<tr>
<td><strong>1030 - 1100</strong></td>
<td>MORNING TEA</td>
</tr>
<tr>
<td><strong>1100 – 1230</strong></td>
<td>Collaboration, Sharing Best Practice and Developing the Virtual Health Information Network (VHIN)</td>
</tr>
<tr>
<td>(Graeme Osborne)</td>
<td>1) Introduction – Tony Blakely (20 mins)</td>
</tr>
<tr>
<td></td>
<td>2) Options for a Charter, or Operating Principles – Simon Ross (30 mins)</td>
</tr>
<tr>
<td></td>
<td>3) Development of themes of activity – Douwes, Mehta, Kvizhinadze, McDonald (20 mins)</td>
</tr>
<tr>
<td></td>
<td>4) Methodological Opportunities – Tony Blakely (10 mins)</td>
</tr>
<tr>
<td></td>
<td>5) <strong>Discussion:</strong> the way forward for VHIN and IDI users (10 mins)</td>
</tr>
<tr>
<td><strong>1230 - 1330</strong></td>
<td>LUNCH</td>
</tr>
<tr>
<td><strong>1330 - 1500 (George Disney)</strong></td>
<td><strong>Challenges, Ground Work Required and Motivation for Using IDI: Examples of Work Underway</strong></td>
</tr>
<tr>
<td></td>
<td>1) Māori research issues – Andrew Sporle (20 mins)</td>
</tr>
<tr>
<td></td>
<td>2) Better Start, National Science Challenge – Rick Audas (15 mins)</td>
</tr>
<tr>
<td></td>
<td>3) Pharmacoepidemiology – Lianne Parkin (15 mins)</td>
</tr>
<tr>
<td></td>
<td>4) The employment and income effects of health conditions – Sylvia Dixon (15 mins)</td>
</tr>
<tr>
<td></td>
<td>5) Family Units in the IDI - Barry Milne (15 mins)</td>
</tr>
<tr>
<td></td>
<td>6) General <strong>Discussion</strong> (10 mins)</td>
</tr>
<tr>
<td><strong>1500 - 1530</strong></td>
<td>AFTERNOON TEA</td>
</tr>
<tr>
<td><strong>1530 – 1700</strong></td>
<td><strong>Existing Epidemiological Data Projects: Lessons Learnt for Future Research</strong></td>
</tr>
<tr>
<td>(Tony Blakely and Simon Ross)</td>
<td>1) PREDICT/VIEW – Daniel Exeter (15 mins)</td>
</tr>
<tr>
<td></td>
<td>2) Ministry of Health analysis using the IDI – first steps, future plans– Steven Johnston (15 mins)</td>
</tr>
<tr>
<td></td>
<td>3) Injury Prevention – Gabrielle Davie (15 mins)</td>
</tr>
<tr>
<td></td>
<td>4) Housing – Michael Baker (15 mnins)</td>
</tr>
<tr>
<td></td>
<td>5) Concluding <strong>Discussion:</strong> main lessons from the day and “what next?” (30 mins)</td>
</tr>
</tbody>
</table>
Introduction to the VHIN

• Context, NZ and Internationally
• Vision
• Options
• Implementation phases, and where we are now (then Simon will take over with where to next)

Sources: my personal involvement in the evolution of the VHIN, and in particular an Options Paper and Implementation Plan prepared by Robin Olds for the VHIN in April 2015 (which is at this Workshop’s website, and is an excellent overview)
Long-standing history of linked data

New Zealand:
  – Using NHI (since c.1990):
    • researchers have led bespoke linkage projects (e.g. PREDICT)
    • ‘thorough’ linkage for planning, analysis and research (e.g. HealthTracker, Virtual Diabetes Register)
  – Using record linkage, researchers have been linking health and other datasets (e.g. NZCMS)

Internationally, bespoke to platforms:
- UK: SHIP and FARR [more 10 care and eHealth Record data than possible in NZ]
- Ontario, Canada: Institute for Clinical Evaluative Sciences
ICES

Health Services (IKN)
- Physician claims
- In-patient discharge abstracts
- Emergency and ambulatory care abstracts
- Prescription drug claims (65 and over)
- Home care claims
- Rehab claims
- Long-term care visits
- In-patient mental health data

People & Geography (IKN)
- People in Ontario since 1985
- Unique individual anonymous # IKN
- Postal Code Conversion/Geographical
- Population Estimates
- Canada Census Profiles
- Death register

Special Collections (IKN)
- Registries: cancer, stroke, CCN, *Birth outcomes
- Federal immigration register
- Corrections
- *First Nation Metis
- Bio-informatics data
- Cell phone records, laboratory data, Clinical trial data
- Developmental Disabilities

Derived chronic conditions (IKN)
(using linked data)
- Diabetes
- Hypertension
- COPD
- CHF
- AMI
- Asthma
- IBD
- Registries – DES/ICD

Real-time (IKN)
- *HOBIC
- *Peritoneal Dialysis

Provider/Facilities
- Physicians
- Hospitals
- Complex care
- Long-term care homes
- Home care

IKN = unique algorithm based on Ontario health card number

Project Dataset

Primary Collected Data And Surveys (IKN)
Moving from bespoke to platform

• Rather than assembling linked data ‘as needed’ after coming up with good research questions, one rather envisages bodies of work (and therefore as yet unarticulated research questions) and links multiple datasets that future researchers can come along to and answer questions with

• But high quality research questions still needed! And (reasonable) quality data... Otherwise we end up in a mess....
Pitfalls to be wary of

GET ALL THE INFORMATION YOU CAN, WE’LL THINK OF A USE FOR IT LATER.

LETS SOLVE THIS PROBLEM BY USING THE BIG DATA NONE OF US HAVE THE SLIGHTEST IDEA WHAT TO DO WITH

© marketoonist.com
NZ Opportunities

• NHI number – ubiquitous national health identifier

• IDI, for many reasons:
  – Creates a data laboratory for access to frequently updated health (only!) data
  – Links health and social sector data – surely some of the best linked data internationally
  – Has the ability going forward to integrate:
    • Richer sub-population data (e.g. Cohort studies, health surveys)
    • e Health Records (recent announcement by Coleman for eHR in 3-5 yrs)

• Our collegial workforce: Govt analysts ↔ academics
VHIN Vision

... to create and sustain an environment that captures value from linking health data collections, through world leading health research, policy development and service planning.
What might the VHIN create?

- Enhanced evidence base to support policy development, improve health and reduce inequalities
- Growth in capacity to create knowledge from large integrated datasets
  - Improved data access and analysis tools
  - People able to ‘do it’, and ‘share it’
  - Code and metadata, shared learnings
  - Derived variables and algorithms
- More efficient delivery of services
- World leading research
Options

1. Support a thought and practice leader to develop an interdisciplinary team with the requisite skills

2. Outsource, e.g. to a consultancy or research team in another country

3. Create a virtual team of network:
   - Network of interdependent interactions sharing a purpose
   - Can be quickly assembled with little infrastructure; nimble
   - Risks, however, about high failure rates and unsustainability

“To be sustainable the [VHIN] must deliver new value... and create win-win situations for all involved, rather than being a burden or resource consuming activity for some so that others can garner the benefits”
VHIN – Implementation stages

• Phase 1: Initiation
  – Organic, through the enthusiasm of key individuals
VHIN – Implementation stages

- Phase 1: Initiation
  - Identify network champions (Osborne, Blakely, Douwes, Milne; Ross)
  - Constitute a core team (VHIN Executive [above + Vandenberg, Jameson, McDonald, Sloane]; Sheree Gibb; funded researchers; project sponsors)
  - Develop a network charter (or “operating principles”) to encourage meaningful participation and avoid free-riding (Simon Ross next session)
  - Develop network culture (today and other activities; reciprocity)
  - Create a platform to support virtual and face-to-face interactions (VHIN FaceBook page; website; ? Use SNZ wiki [in development, for code, etc])
  - Deliver some demonstration projects (Next slide, and later sessions)
  - Establishment funding (Catalyst projects by MoH [in kind], Auckland/Otago/Massey universities; Healthier Lives NSC; enormous good-will and ‘in-kind’)
VHIN Demonstration Projects

Catalyst projects, funded by Ministry, Otago, Massey and Auckland Universities:

- Core staff: Sheree Gibb, Simon Ross (June Atkinson); Future model?
- Getting the denominator right (Host – Uni Auckland; Champion - Jackson)
- Cost of CVD in New Zealand (Host – Otago Uni; Champion – Blakely)
- Occupational and Pharmaceutical risk factors for Congenital Malformations (Host – Massey Uni; Champions: Mannetje, Borman, Eng, Douwes)
- Enhancing the Virtual Diabetes Register (Host- Ministry of Health; Champions: Evison)
- Improving the Health Tracker (Host Ministry of Health; Champion: Fawcett)

Healthier Lives National Science Challenge:

- Moving beyond ‘demonstration’
- Will be explained by Andrea McDonald later
VHIN – Implementation stages

• Phase 1: Initiation
• Phase 2: Consolidation
• Phase 3: Expansion
  – Exactly how pending, but possibly including:
    • Recruitment of members, develop network structure, regular events, demonstrating value, developing capacity and team of analysts ‘for hire’, partnerships
    • Depends on operating principles (aka charter), funding arrangements, culture that evolves – Simon’s talk next
• Phase 4: Exerting Influence
Key obstacle to overcome: free-rider
The free-rider problem...

• ... occurs when those who benefit from resources, goods, or services do not pay for them, which results in an under-provision of those goods or services. ([https://en.wikipedia.org/wiki/Free_rider_problem](https://en.wikipedia.org/wiki/Free_rider_problem))

• ... is common among public goods. These are goods that have two characteristics: non-excludability — non-paying consumers cannot be prevented from using it — and non-rivalry — when you consume the good, it does not reduce the amount available to others. The potential for free riding exists when people are asked to voluntarily pay for a public good. ([https://en.wikipedia.org/wiki/Free_rider_problem](https://en.wikipedia.org/wiki/Free_rider_problem))

• Assuming the VHIN produces ‘things’ useful to the next wave of researchers, how do we avoid the free-rider problem? Or put another way, make the network sustainable?
We have already generated health system costs leading up to and following diagnosis of cancers [Blakely et al. Patterns of Cancer Care Costs in a Country With Detailed Individual Data. Medical Care 2015;53(4):302-09.]

Uses HealthTracker data, but outside of IDI

Goal of this project:
- Repeat exercise for CVD diagnoses
- Leave derived cost data and code ‘behind’ in the VHIN for other researchers to use
Example from cancer

Predicted Excess Cost per month with 95% confidence intervals
Bit more on CVD health system costs

• Will be done for multiple ICD groupings, e.g.:
  – Myocardial infaction
  – Angina
  – Stroke (haemorrhagic versus ischaemic if possible)

• First catalyst project to be undertaken.
  – Will be ‘championed’ by Blakely
  – And undertaken by Sheree Gibb (employed by VHIN for 6 months, located in the Ministry, supervised by Simon Ross and Tony Blakely) and Giorgi Kvizhinadze (BODE³ researcher)
Methodological opportunities

• Key point: Lots of ‘good’ will come from simple and descriptive analyses of linked data for NZ policy and local consumption. But we should aim to also produce novel knowledge internationally, marrying great data with thorough and innovative methods.

• To cover briefly in next few minutes (just some topics!):
  – Data curation and management (e.g. imputing)
  – Addressing mismeasurement
  – Causal inference methods
  – Academic centre(s) or collaborations of excellence required
Data management and curation

- SQL and SAS expertise
- Probabilistic record linkage, and the right linkage for the right denominator for each project
- Feeding back to sector findings and missing data problems, to increase quality of data entering IDI
Missing data

• There is going to be lots of it!

• For longitudinal analyses, could quickly lead to high percentage of observations being discarded, and then potentially (especially in the eye of journal referees!) selection bias
  – i.e.: the association of X with Y you observe not necessarily be the same as that in the total eligible population

• Imputation for missing data ‘mainstream’ now, but difficult for:
  – Nominal variables (e.g. ethnicity)
  – Longitudinal data

• Ideally, we want increased statistical expertise
Measurement error

• Rife! (But we like to pretend it does not exist....)

• Correcting for measurement error can substantively change findings, e.g.:
  – residual confounding
  – exposure misclassification

• Expertise in addressing measurement error in NZ could/should be deepened, e.g.:
  – Econometric techniques
  – Regression calibration
  – Multiple over-imputation, and interesting merging with imputation
  – Simpler bias analysis methods if appropriate
Causal inference ‘stable’ of methods

• There has been a flourishing of methodological approaches to ‘try’ and estimate causal effects, leveraging off:
  — Counterfactual models and potential outcomes approach
  — DAG (directed acyclic graphs) clarity

• Examples:
  — Marginal structural models and inverse probability of treatment weighting
    • Useful when variable effected by past exposure confounds future exposure-outcome association
    • Useful for questions like “How much of the association of ethnicity with CVD incidence is due to mediating risk factors like smoking and health services?”
    • Complex biases to address, e.g. collider biases, exposure-induced mediator outcome confounding
Contention: ↑ academic excellence needed (or highly desirable)

• The IDI (and VHIN) will work for local analyses and policy
• But to really do well, we need to ingeniously scale up academic expertise and leadership in biostatistics and ‘big data’, e.g.:
  – More Assoc Profs in statistics using longitudinal data
  – Programmes and centres of research excellence
• This will help lead to IDI analyses being in high impact journals internationally, sustaining funding and world-leading expertise
• Approaches: creating research centre(s) of excellence, VHIN, international collaborations, dispersed academic leadership
• What do you think?