Biases arising from using linked administrative data for research

Richard Shaw
Research Associate
University of Glasgow
Motivation - Evaluating social polices using data from the 100 Million Brazilian Cohort

- Family members who applied for social assistance (2001 to 2015) and registered in the Cadastro Único para Programas Sociais (CadÚnico) database
  - 114 Million people - 57% those most disadvantaged and in need of support
- Social policy data include:
  - Receipt of Bolsa Familia and Minha Casa, Minha Vida
  - Socio-demographic information
- Linked to health related data including:
  - Mortality / Sistema de Informação sobre Mortalidade (SIM)
  - Hospital care / Sistema de informações Hospitalares do SUS
- Early results inconsistent with theory:
  - Lack of appreciation of how administrative data differs from research data?
- **Aim**: Improve understanding of potential sources of bias when using linked administrative data in research
A schematic illustrating with the 100 Million cohort possible sources of bias created by linking administrative data.
Registration and recording (Denominator)

- Recording and registration will lead to different types of biases depending on role data is used for in study
- Registration in the CadÚnico principally determines the denominator of the population
  - Selection bias – representativeness and generalisability of sample
  - Different implication for descriptive and causal epidemiology
- Select areas or groups with fewer registration problems to improve internal validity at the expense of external validity
Registration and recording (Numerator)

- The non registration or recording health events impact on the numerator
  - Misclassification bias (always underestimate)
  - Potential to alter social gradients

- Mortality registration by education (Costa 2020)
  - 98.6% registered in most educated municipalities
  - 93.8% registered in least educated municipalities

- Health-care funded by the Unified Health Care system (Sistema Unico de Saude) is registered
  - For elective services many patients that can afford it use private healthcare services
  - For some health outcomes that could lead to inflated estimates of health inequalities

Linkage

- The linkage process 100M cohort and Mortality Information System (SIM) risks errors (misclassification biases)
  - Missed link - a person may be considered alive when not
  - False link – a person may classified as dead when alive
- At individual level missed links very hard to identify
  - Require complete population coverage to identify
  - Small proportion of 100 Million cohort expected to die in any period
  - Not all deaths in SIM will be for the 100 Million cohort
- Identifying missed links and false links at individual level may not be possible with anonymized data
- Area level information can be used to identify areas where linkages may be more accurate
Cleaning and Coding

- Two groups with different skills and knowledge
- Data scientists linking data together
  - Clean and code linkage variables
  - Documenting distribution of linkage variable could identify source of bias
  - Requires planning in advance if linkage is computationally intensive greater collaboration between data scientists and analysts required
- Analysts - Epidemiologists, statisticians, demographers
  - Cleaning and coding should be informed by knowledge of the whole process
Options to help minimise potential sources bias

- Consider the whole pathway deriving linked data for analysis
- Identify and plan for possible sources of bias at start of project
- Collaboration needed between analysts and data scientists linking the data
- Ability to identify potential sources of bias needs to be considered alongside issues of confidentiality and privacy
- Training needed on sources of bias for all stakeholders involved in using administrative data for research
- Focusing on areas or groups with lower risk of bias to improve internal validity
Acknowledgements

Collaborators

• Julia Pescarini, Elzo Junior, Andressa Siroky, Robespierre Pita, Mauricio Barreto - CIDACS, Salvador
• Mirjam Allik, Desmond Campbell, Ruth Dundas, Alastair Leyland, Srinivasa Vittal Katikireddi - University of Glasgow
• Katie Harron, University College London

Funders

• NIHR Public Health Research Board

richard.shaw@glasgow.ac.uk  @rickwahs