IMPACT OF LINKAGE ERRORS ON EPIDEMIOLOGICAL ANALYSES

Big Data Bahia 2018

Elizabeth Williamson

London School of Hygiene & Tropical Medicine, UK

Health Data Research UK

CENTRE for STATISTICAL METHODOLOGY

LONDON SCHOOL of HYGIENE & TROPICAL MEDICINE

Overview

- Evaluating and dealing with the impact of linkage errors has been identified as a priority for researchers in this area (e.g. Jorm, 2015)

- Impact of errors
- Ways to account for linkage error

Impact of linkage errors

Will depend on

- The analysis of interest
- The structure of the data and role of the linkage
- The extent and type of linkage errors

Analysis

- Focus on simple epidemiological analysis

- Prevalence or incidence of an event (e.g. mortality, cancer diagnosis)
- Comparison of events between groups

Purpose of linkage

- Adding outcome information
  - E.g. linking to mortality data to determine vital status
  - Linked = event
  - Unlinked = no event

Misclassification

- Defining study population
  - E.g. cohort participants diagnosed with cancer, by linking to cancer registry
  - Linked = included in analysis
  - Unlinked = excluded

Selection bias

- Adding covariate information
  - E.g. detailed measures of socioeconomic status, by linking to social care data
  - Linked = extra data
  - Unlinked = missing data (potentially excluded)

Missing data

Type of errors

Truth

Link status

	Match (Record from same person)	Non-match (Records from different individuals)
Link	Identified match	**False match**
Non-link	**Missed match**	Identified non-match

IMPACT

OF LINKAGE ERRORS

Linking to event data: Estimating prevalence or incidence

- Missed matches
  - Underestimate prevalence / incidence
- False matches
  - Overestimate prevalence / incidence
  - Overestimation is inversely related to true prevalence (Brenner, 1997)
    - bigger errors for when prevalence is small
    - rare conditions are worse affected by false matches

Linking to event data: Comparing groups

Standard misclassification scenario:

- Non-differential
  - Linkage errors same in groups being compared
    - Same proportion of false matches and missed matches across groups
  - Moves estimates towards null, i.e. dilutes estimates of effect
  - Estimates generally fairly robust to non-differential missed matches
- Differential
  - Linkage errors different in groups being compared
  - Can cause bias in either direction

Example 1: Mortality by ethnicity in US

Background:

- Hispanics have been found to have better mortality than non-Hispanic whites in the US
- This is at odds with the expected effect of socioeconomic status
- Linkage is likely to be worse for Hispanic people

- Nationally representative cohort (National Health Interview Survey) in the US
- Linked to national cause of death data (National Death Index)
- Probabilistic linkage:
  - Split into classes (which characteristics match)
  - Thresholds of match scores within classes determine links

Reference Lariscy JT. *J Aging Health*, 2011

Example 1: Mortality by ethnicity in US

HAZARD RATIOS FOR MORTALITY	Relaxed	Usual thresholds	Tighter
US born, non-Hispanic white	Ref	Ref	Ref
Foreign born, non-Hispanic white	0.81	0.78	0.77
US born, Hispanic	1.14	1.10	1.06
Foreign born, Hispanic	1.24	0.97	0.78

more deaths
more false matches
fewer missed matches

fewer deaths
fewer false matches
more missed matches

Reference Lariscy JT. *J Aging Health*, 2011

Linking to define study population

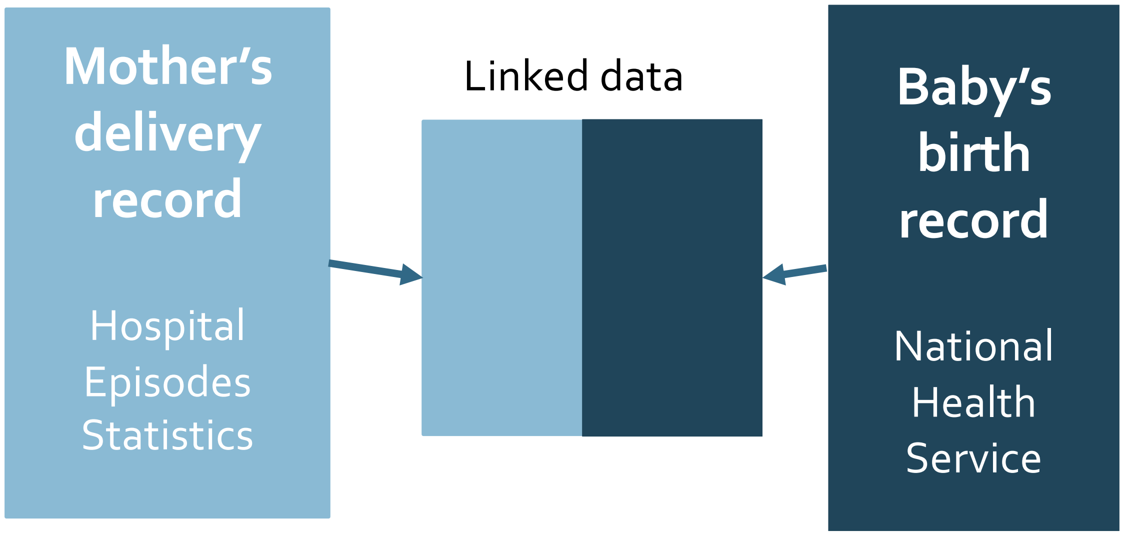
- Linkage defines who is included/excluded
  - E.g. Analyse only linked records

- Missed matches
  - Lower sample size
  - Potential selection bias

- False matches
  - Inclusion of irrelevant people / units of analysis
  - Noise, dilution of effects
  - Potential bias

Example 2: Mother-baby cohort

- Mother-baby cohort
- 42% of baby records linked using deterministic linkage
- 98% linked using probabilistic linkage
- Also had subset of “gold-standard” linkage



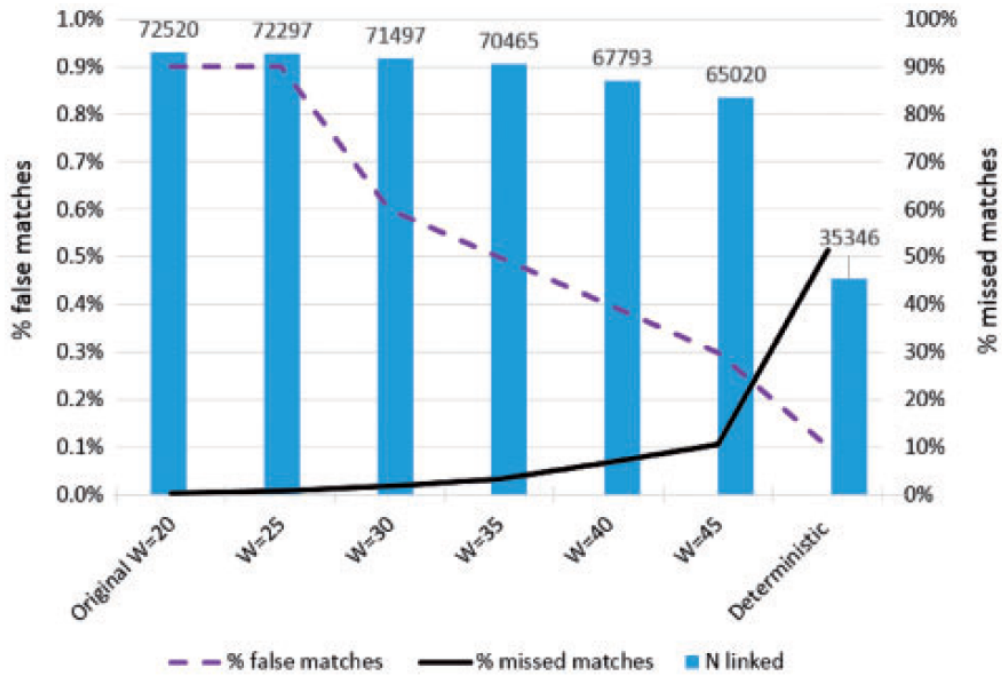
Reference: Harron K et al. *Int J Epi*, 2017

Example 2: Mother-baby cohort

- Power issues
- Loss of sample size

More false matches

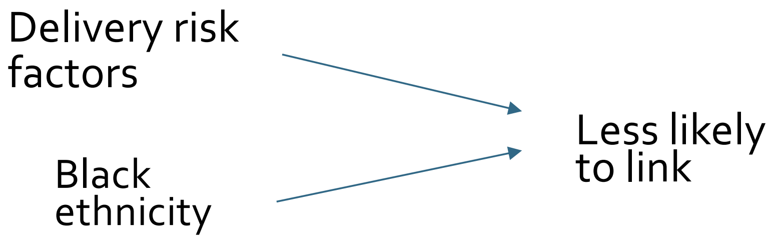
More missed matches



Reference: Harron K et al. *Int J Epi*, 2017

Example 2: Mother-baby cohort

- Association between:
  - Black ethnicity (exposure) and
  - Having risk factors for delivery (outcome)



- If both these factors affect the probability of being linked

  ...i.e. affect the probability of being included in the analysis

  ...then selection bias can occur

- Gold standard:
  - 6.5% of mothers with delivery risk factors are black
- Deterministic:
  - 4.7% of mothers with delivery risk factors are black

Reference: Harron K et al. *Int J Epi*, 2017

Example 2: Mother-baby cohort

	Gold standard	Probabilistic	Tighter	Deterministic
Pre-term birth (%)	7.65%	7.64%	7.31%	7.43%
Black vs white ethnicity:				
OR (delivery risk factors)	0.98	0.97	0.89	0.80
95% CI	(0.88, 1.09)	(0.87, 1.08)	(0.79, 1.01)	(0.66, 0.96)

Reference: Harron K et al. *Int J Epi*, 2017

METHODS TO HANDLE

LINKAGE ERRORS

Approaches to handling linkage errors

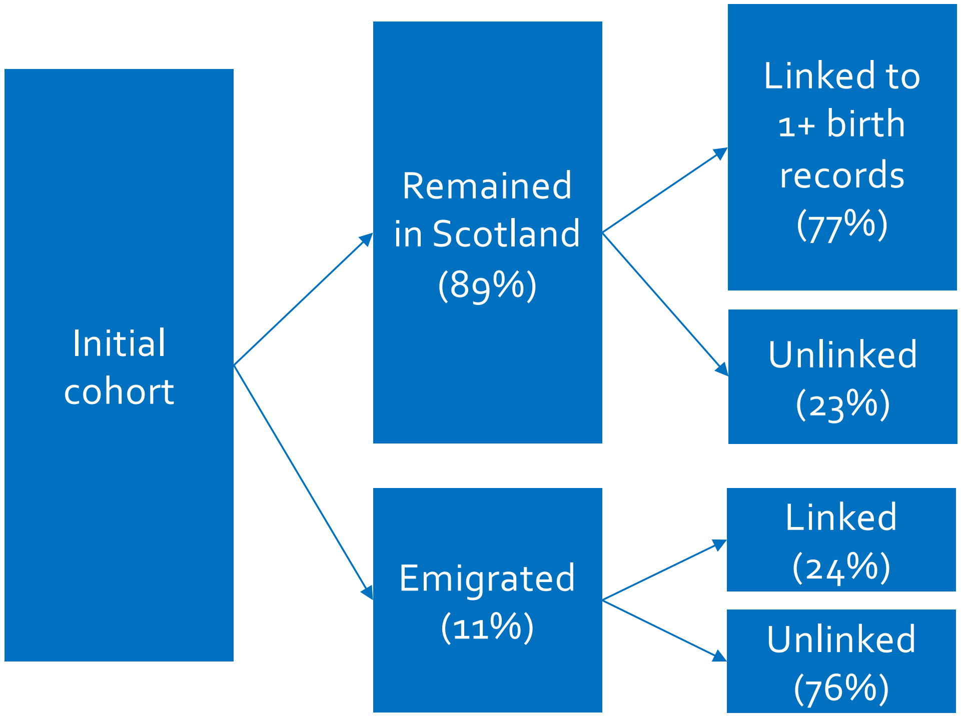
- Ignore the linkage error
- Quantify the bias (sensitivity analysis)
  - E.g. changing linkage thresholds
  - E.g. exploring mechanisms of linkage error
- Correcting for the bias and incorporating linkage uncertainty
  - Treat as missing data problem

Example 3: Childhood SES and childbirth

- Aim: To assess the effect of childhood socio-economic (SES) status on likelihood of having children
- Children of the 1950's cohort sub-study
  - 4997 women in Aberdeen (Scotland) with perinatal and childhood data
- Linked to:
  - Scottish Maternity Record (incomplete until 1976)
  - AMND (Aberdeen only)

Reference: Nitsch et al. *JRSS A*, 2006

Example 3: Childhood SES and childbirth

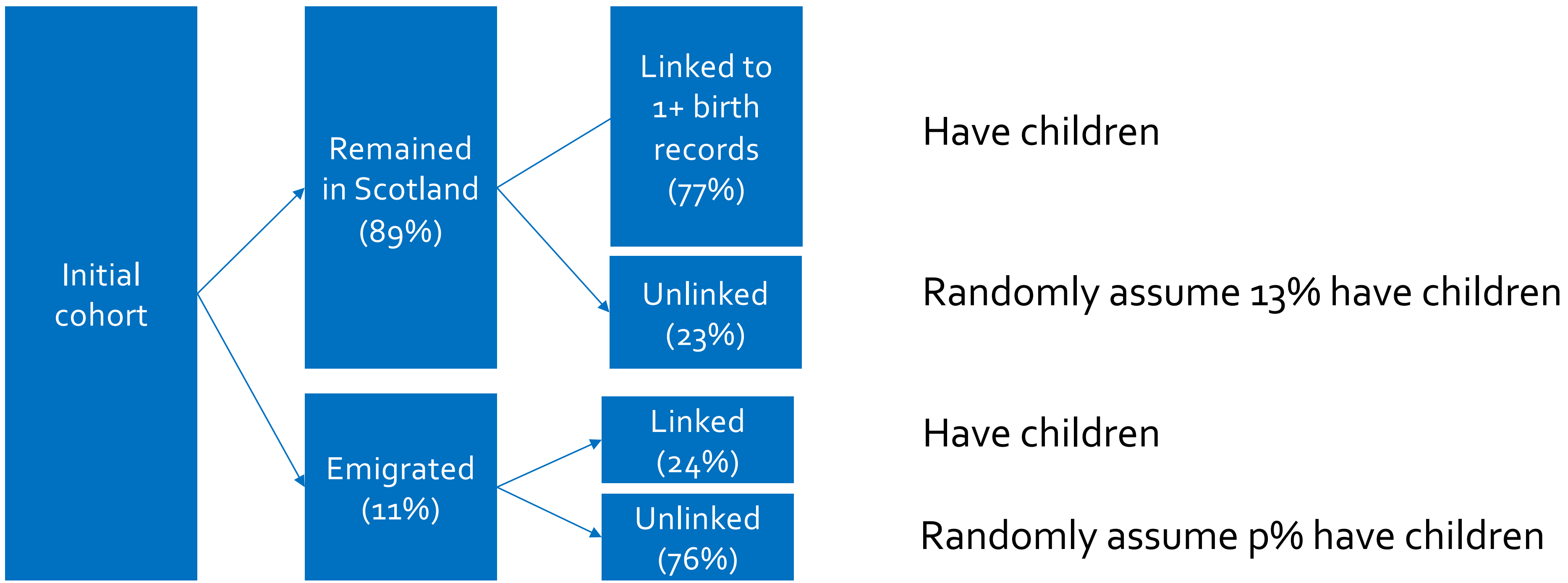


- Using “gold-standard” data
- Estimated:
  - Specificity: 97%
  - Sensitivity: 87%

  i.e. may have missed 13% births

- Better IQ
- Better growth
- Higher socioeconomic status

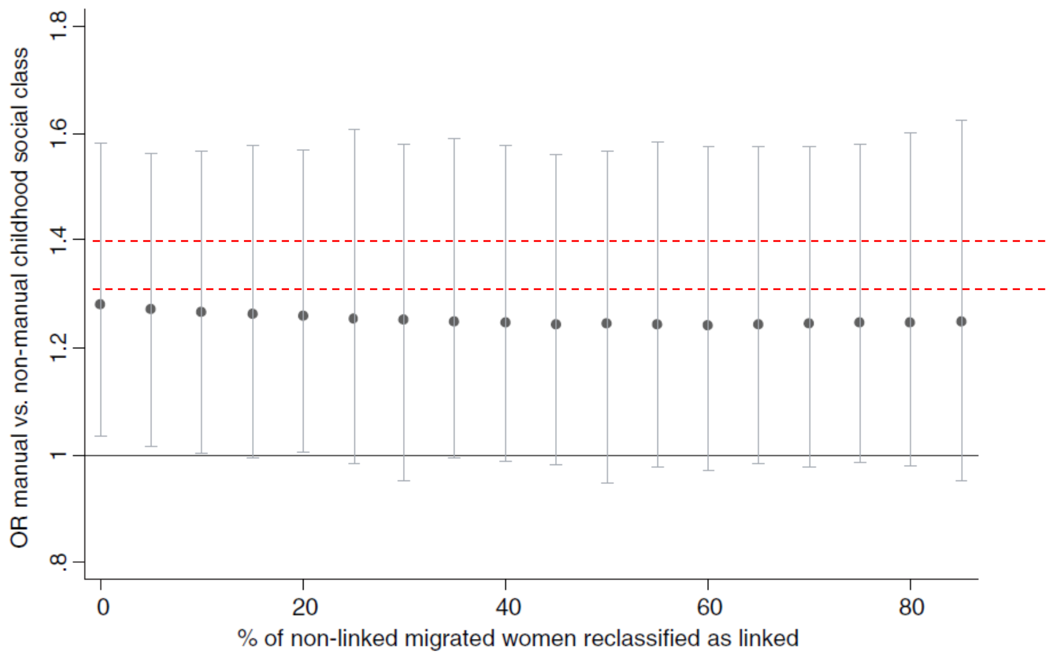
Reference Nitsch et al. *JRSS A*, 2006



- For each scenario, simulated 1000 datasets (with re-assignment)
- Calculated estimated OR for socioeconomic status (manual vs non-manual)
- Estimated 95% CI was minimum bound – maximum bound over 1000 datasets.

Reference: Nitsch et al. *JRSS A*, 2006

Example 3: Childhood SES and childbirth



Original linked data
Adjusting for migration status

- Estimated effect of childhood SES on childbirth was robust to misclassified status of migrants

Reference: Nitsch et al. *JRSS A*, 2006

Carrying through uncertainty

- Various methods proposed by Goldstein et al and colleagues
- Essentially, rephrase aim
  - Not to link data
  - But to add information about particular **variables of interest** to analysis dataset
- Recasts the problem as a missing data problem

Sex	Age	Height	Dead
M	35	1.66	?
F	82	1.43	?
F	79	1.62	?
M	56	1.82	?

Carrying through uncertainty

- E.g. Suppose we are linking a cohort data to mortality records (so link = dead)

Sex	Age	Height	Dead	
M	35	1.66	1	Certain links
F	82	1.43	?	Uncertain links
F	79	1.62	?	
M	56	1.82	0	Certain non-links

- Apply multiple imputation
- Can incorporate:
  - Outcomes in potential links (where multiple)
  - Match probabilities / weights

CONCLUSIONS

Conclusions

- Impact of linkage errors depends on structure of data, the role of the linkage, the analysis, and the extent and type of errors
- Substantial bias can occur, particularly for comparisons involving groups with particularly poor linkage

- Methods to handle linkage error include:
  - Sensitivity analysis
  - Imputation-based approaches

References

- Jorm L. Routinely collected data as a strategic resource for research: priorities for methods and workforce. Public Health Res Pract, 2015:25;e2451540.
- Brenner H, Schmidtmann I, Stegmaier C. Effects of record linkage errors on registry-based follow-up studies. Stats in Med, 1997, 16:2633-2643.
- O'Reilly D, Rosato M, Connolly S. Unlinked vital events in census-based longitudinal studies can bias subsequent analysis. *J Clin Epi*, 2008, 61(4):380-385.
- Lariscy JT, Differential record linkage by Hispanic ethnicity and age in linked mortality studies. J Aging Health, 2011, 23(8):1263-1284.
- Harron K, Doidge J, et al. A guide to evaluating linkage quality for the analysis of linked data, Int J Epi, 2017, 1699-1710.
- Nitsch D, deStavola B, et al. Linkage bias in estimating the association between childhood exposures and propensity to become a mother: an example of simple sensitivity analysis. JRSS A, 2006, 169(3):493-505.
- Goldstein H, Harron K, Wade A. The analysis of record linked data using multiple imputation with data value priors. Stat in Med, 2012.