The Fundamental Role of Linkage Uncertainty in the Epidemiological Analysis of Big Data

Evelyn Lauren[1,2], Dorina Onoya[1], Koleka Mlisana[3], Jacob Bor[1,4,5]

[1]Health Economics and Epidemiology Research Office, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa
[2]Department of Biostatistics, Boston University School of Public Health, Boston, MA, USA
[3]National Health Laboratory Service, Johannesburg, South Africa
[4]Department of Global Health, Boston University School of Public Health, Boston, MA, USA
[5]Department of Epidemiology, Boston University School of Public Health, Boston, United States

Health Economics and Epidemiology Research Office
HE²RO
Wits Health Consortium
University of the Witwatersrand

BOSTON UNIVERSITY

Fundamental role of linkage error in epidemiology

- Epidemiologists increasingly use linked administrative data

	Venue: Ballroom East
16h30 - 18h30 | Interactive sessions	**INT01: Are traditional cohorts outdated?**
	Chair: Brigid Lynch (Cancer Council Victoria, Australia)
	Mauricio Lima Barreto (Center of Data and Knowledge Integration for Health, Brazil)
	Karen Canfell (The Daffodil Centre, Cancer Council NSW / University of Sydney, Australia)
	Naja Hulvej Rod (University of Copenhagen, Denmark)
	Yuan Lin (Nanjing Medical University, China)
18h30 - 20h30 | Symposia	**Venue: Meeting Block 1.61-1.64**
	SYM09: Population-based epidemiology in the era of data science and routine health data
	Andrew Boulle (University of Cape Town, South Africa)
	Spiros Denaxas (University College London, United Kingdom)
	Maurício Lima Barreto (Center of Data and Knowledge Integration for Health, Brazil)

Fundamental role of linkage error in epidemiology

- Epidemiologists increasingly use linked administrative data
- Probabilistic record linkage (e.g. Fellegi-Sunter) tolerates linkage error
- Extent of linkage error in data is rarely reported in epidemiological analyses
- Impact of linkage error on bias and variance of estimates rarely estimated

Fundamental role of linkage error in epidemiology

- Epidemiologists increasingly use linked administrative data
- Probabilistic record linkage (e.g. Fellegi-Sunter) tolerates linkage error
- Extent of linkage error in data is rarely reported in epidemiological analyses
- Impact of linkage error on bias and variance of estimates rarely reported
- No textbook methods for estimating bias/variance due to linkage error
- **This is an under-studied, first-order problem: in complete population analyses, there is ZERO sampling error; but there is still linkage error.**

Linkage error is a distinct source of uncertainty

Sampling Error: $\hat{\beta} \rightarrow N(\beta, \frac{\sigma^2}{n})$ Central Limit Theorem

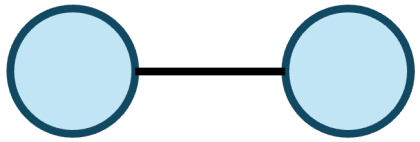
- The basis for statistical inference in large samples; e.g. 95% CI's

Linkage Error: $\hat{\beta} \rightarrow ?\,(?\,,?\,)$ Unknown Asymptotic Distribution

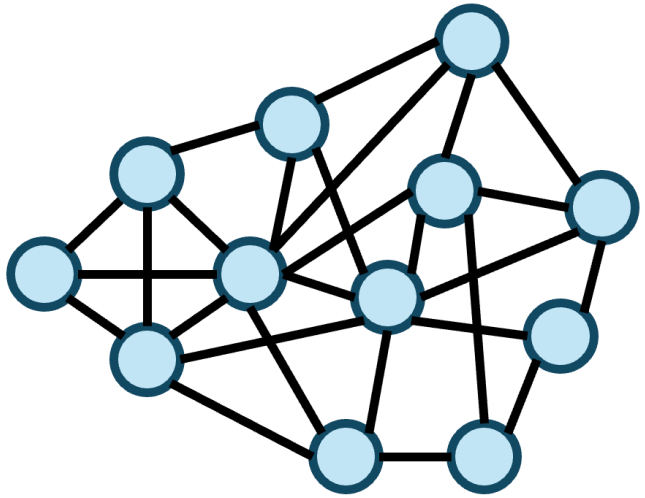
- What is the bias and the variance of $\hat{\beta}$ due to linkage error?

Linkage as a network problem

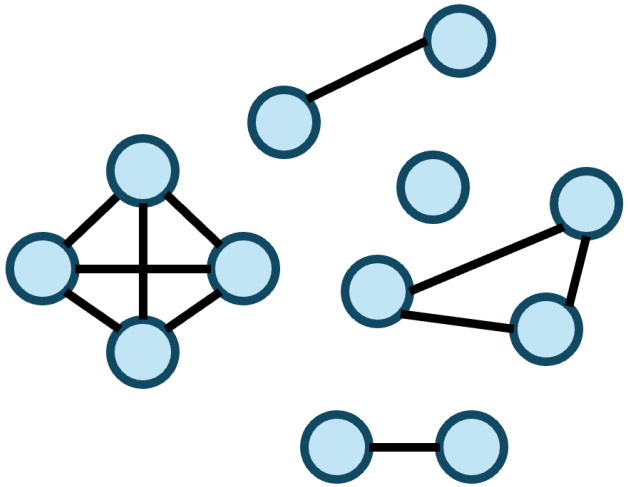
Data entry errors can lead to different representations of individuals



Assess similarity of record pairs based on identifying characteristics



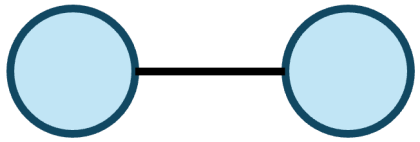
These comparisons form a network structure



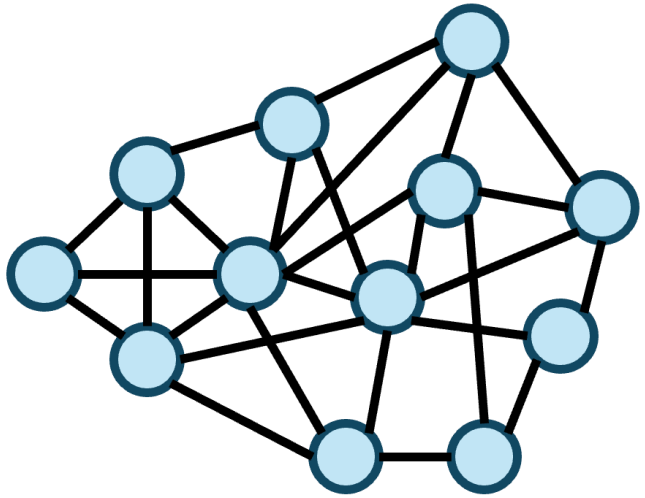
Task of record linkage: identify clusters belonging to underlying individuals

Linkage as a network problem

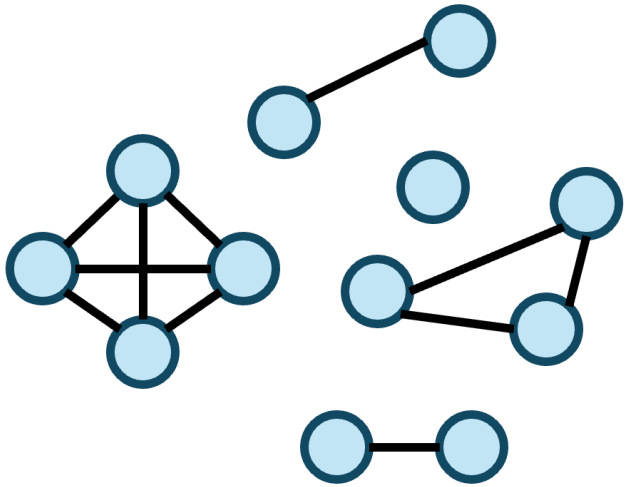
Data entry errors can lead to different representations of individuals



Assess similarity of record pairs based on identifying characteristics



These comparisons form a network structure

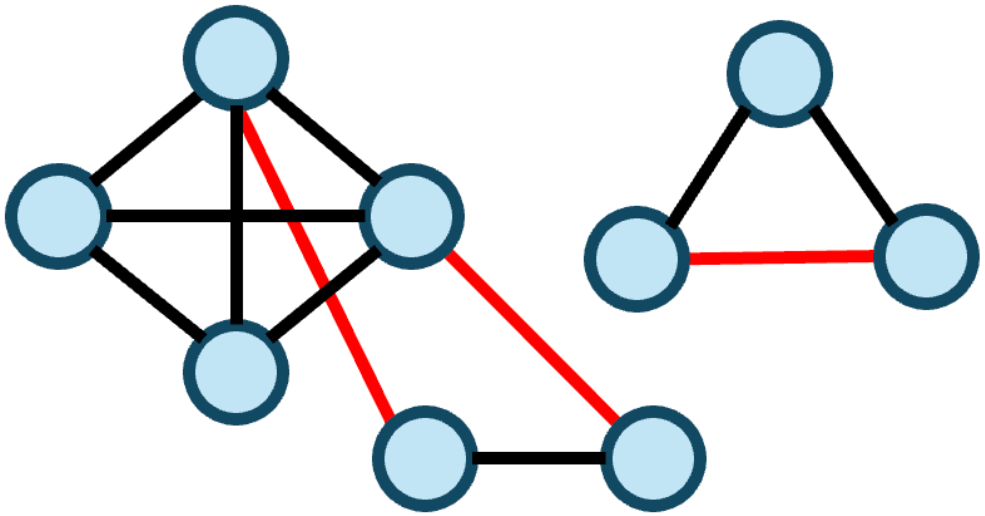


Task of record linkage: identify clusters belonging to underlying individuals

Impact of linkage error depends on network structure of the linkage problem

Type 1 Linkage Error: Overmatching

Overmatching (1-PPV): Falsely link records that belong to different individuals



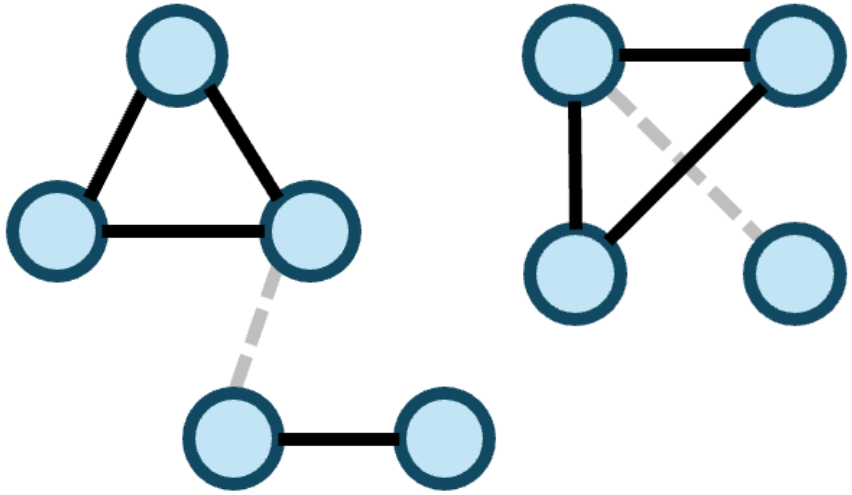
Truth: 3 individuals with 4, 2, 3 records
Observed: 2 individuals with 6, 3 records

Implications of overmatching:

- Individuals have additional false data points
- Individuals that should have existed are now missing

Type 2 Linkage Error: Undermatching

Undermatching (1-SEN): Failing to link records belonging to an individual



Truth: 2 individuals with 5, 4 records
Observed: 4 individuals with 3, 2, 3, 1 records

Implications of undermatching:

- Individuals have missing data points
- We create additional false individuals

How does linkage error affect bias and variance of point estimates?

Methods

- Simulation using network structure of a "real-world" linkage dataset
- Assess implications for bias and variance of point estimates for:
  - Different levels of linkage error (Sen/PPV combinations)
  - Different sample sizes (N)
  - Different analyses (cross-sectional, longitudinal, regression, prediction)

Dataset: South African NHLS National HIV Cohort

- All laboratory records of all patients seeking care in public sector facilities
- Individuals may have multiple lab records; lots of typographical errors
- Linkage algorithm created a unique patient ID with 94% SEN, 99% PPV
- This analysis used completely de-identified data on network structure
- We simulated datasets with different SEN / PPV combinations by randomly introducing linkage errors

BMJ Open

Cohort profile: the South African National Health Laboratory Service (NHLS) National HIV Cohort

William B MacLeod,[1,2] Jacob Bor,[2,3] Sue Candy,[4] Mhairi Maskew,[2] Matthew P Fox,[2,3] Katia Bulekova,[5] Alana T Brennan,[2,3] James Potter,[1] Cornelius Nattey,[2] Dorina Onoya,[2] Koleka Mlisana,[6] Wendy Stevens,[7,8] Sergio Carmona[7]

NATIONAL HEALTH LABORATORY SERVICE

Results

How do linkage errors affect the **bias** of point estimates?

Cross-sectional count

Note:
- Green indicates overestimation
- Red indicates underestimation
- Darker shade indicates bigger deviation from the true outcome

N of patients entering HIV care between 2012-2016

		SEN					
		100	99	95	90	85	80
PPV	100	**14,393**	14,590 (+1.4%)	15,439 (+7.3%)	16,487 (+14.5%)	17,514 (+21.7%)	18,485 (+28.4%)
	99	14,206 (-1.3%)	14,401 (+0.1%)	15,238 (+5.9%)	16,311 (+13.3%)		
	95	13,353 (-7.2%)	13,557 (-5.8%)	14,429 (+0.3%)	15,525 (+7.9%)		
	90	12,025 (-16.5%)	12,212 (-15.2%)	13,109 (-8.9%)	14,290 (-0.7%)		
	85	10,418 (-27.6%)				14,014 (-2.6%)	
	80	8,534 (-40.7%)					13,556 (-5.8%)

- Lower SEN (undermatching) overestimates outcome
- Lower PPV (overmatching) underestimates outcome
- When PPV ≈ SEN, bias due to linkage error is small

Longitudinal proportion

Note:
- Green indicates overestimation
- Red indicates underestimation
- Darker shade indicates bigger deviation from the true outcome

24-month retention (%) among patients entering care

		SEN					
		100	99	95	90	85	80
PPV	100	**38.7**	38.3 (-1.0%)	36.7 (-5.2%)	35.1 (-9.3%)	33.7 (-12.9%)	32.7 (-15.5%)
	99	38.9 (+0.5%)	38.5 (-0.5%)	36.9 (-4.6%)	35.2 (-9.0%)		
	95	40.0 (+3.4%)	39.5 (+2.3%)	37.8 (-2.3%)	35.9 (-7.1%)		
	90	42.3 (+9.3%)	41.7 (+7.9%)	39.6 (+2.5%)	37.3 (-3.6%)		
	85	45.9 (+18.6%)				37.3 (-3.6%)	
	80	52.3 (+35.1%)					37.9 (-2.1%)

- Lower SEN (undermatching) <u>underestimates</u> outcome
- Lower PPV (overmatching) <u>overestimates</u> outcome
- When PPV ≈ SEN, bias due to linkage error is small

Regression coefficient

Note:
- Green indicates overestimation
- Red indicates underestimation
- Darker shade indicates bigger deviation from the true outcome

Risk ratio (RR) of high vs. low income on 24-month retention

		SEN					
		100	99	95	90	85	80
PPV	100	**3.72**	3.67 (-1.3%)	3.48 (-6.5%)	3.27 (-12.1%)	3.10 (-16.7%)	2.95 (-20.7%)
	99	3.67 (-1.3%)	3.62 (-2.7%)	3.45 (-7.3%)	3.25 (-12.6%)		
	95	3.45 (-7.3%)	3.40 (-8.6%)	3.24 (-12.9%)	3.06 (-17.7%)		
	90	3.11 (-16.4%)	3.06 (-17.7%)	2.91 (-21.8%)	2.77 (-25.5%)		
	85	2.71 (-27.2%)				2.35 (-36.8%)	
	80	2.24 (-39.8%)					1.98 (-46.8%)

- Income is a simulated exposure
- Regression estimate strongly attenuated towards null
- Even when PPV ≈ SEN

Prediction model

Note:
- Green indicates overestimation
- Red indicates underestimation
- Darker shade indicates bigger deviation from the true outcome

AUC of predicted 24-month retention (based on age, income)

		SEN					
		100	99	95	90	85	80
PPV	100	**0.763**	0.759	0.745	0.730	0.718	0.708
	99	0.762	0.758	0.744	0.729		
	95	0.757	0.753	0.738	0.723		
	90	0.747	0.742	0.727	0.711		
	85	0.735				0.685	
	80	0.716					0.655

- Linkage errors lower prediction performance
- Lower PPV does not offset lower SEN

Results

How does linkage error affect the **variance** of point estimates?

How does variance change with PPV/SEN?

Estimand: % retained in care in 24 months (%)

Standard deviation of estimate due to linkage error (80 simulated datasets)							
		SEN					
		100	99	95	90	85	80
PPV	100	**0**	0.05	0.15	0.15	0.12	0.03
	99	0.04	0.07	0.11	0.15		
	95	0.08	0.15	0.15	0.21		
	90	0.18	0.18	0.28	0.13		
	85	0.22				0.13	
	80	0.37					0.27

Variance due to linkage error increases as SEN and PPV decrease

How does variance change with sample size?

Estimand: % retained in care in 24 months (90% PPV)

	Standard deviation of estimate due to:	
Sample size (N=12,025)	**Linkage error**	**Sampling error**
1/8 * N	0.53	1.30
1/4 * N	0.38	0.91
1/2 * N	0.24	0.64
N	**0.18**	**0.45**
2 * N	0.13	0.32
4 * N	0.09	0.23
8 * N	0.06	0.16
$\sigma/\sqrt{N}$	$19.7/\sqrt{N}$	$49.3/\sqrt{N}$

- **Conventional SEs are 25% too small**
- Variance due to linkage error declines approximately $\propto 1/\sqrt{N}$
- Ratio of linkage error to sampling error may vary by analysis and data.

Conclusion

Linkage error can lead to substantial bias in point estimates

- Bias depends on linkage PPV and SEN; however, studies don't always report
- When PPV ≈ SEN, bias is minimal for cross-sectional and longitudinal point estimates; for regression and prediction, linkage error is like misclassification
- Further research could develop approaches to adjust for linkage error

Linkage error leads to added variance in point estimates

- Conventional standard errors are perhaps 25% too small
- However, variance due both sampling and linkage error is low when N is large

Linkage error has important and predictable impacts on bias and variance.

These impacts can be estimated, should be transparently reported, and adjusted for in analyses.

Thank you

This is work in progress. Your feedback will make it better!

Jacob Bor, jbor@bu.edu **Evelyn Lauren, elauren@bu.edu**

Health Economics and Epidemiology Research Office
HE²RO
Wits Health Consortium
University of the Witwatersrand

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

BOSTON UNIVERSITY

NATIONAL HEALTH LABORATORY SERVICE

health
Department: Health
REPUBLIC OF SOUTH AFRICA

NIH National Institute of Allergy and Infectious Diseases