The Farr Institute of Health Informatics Research

Evaluating bias due to linkage error in anonymised linked data

Katie Harron

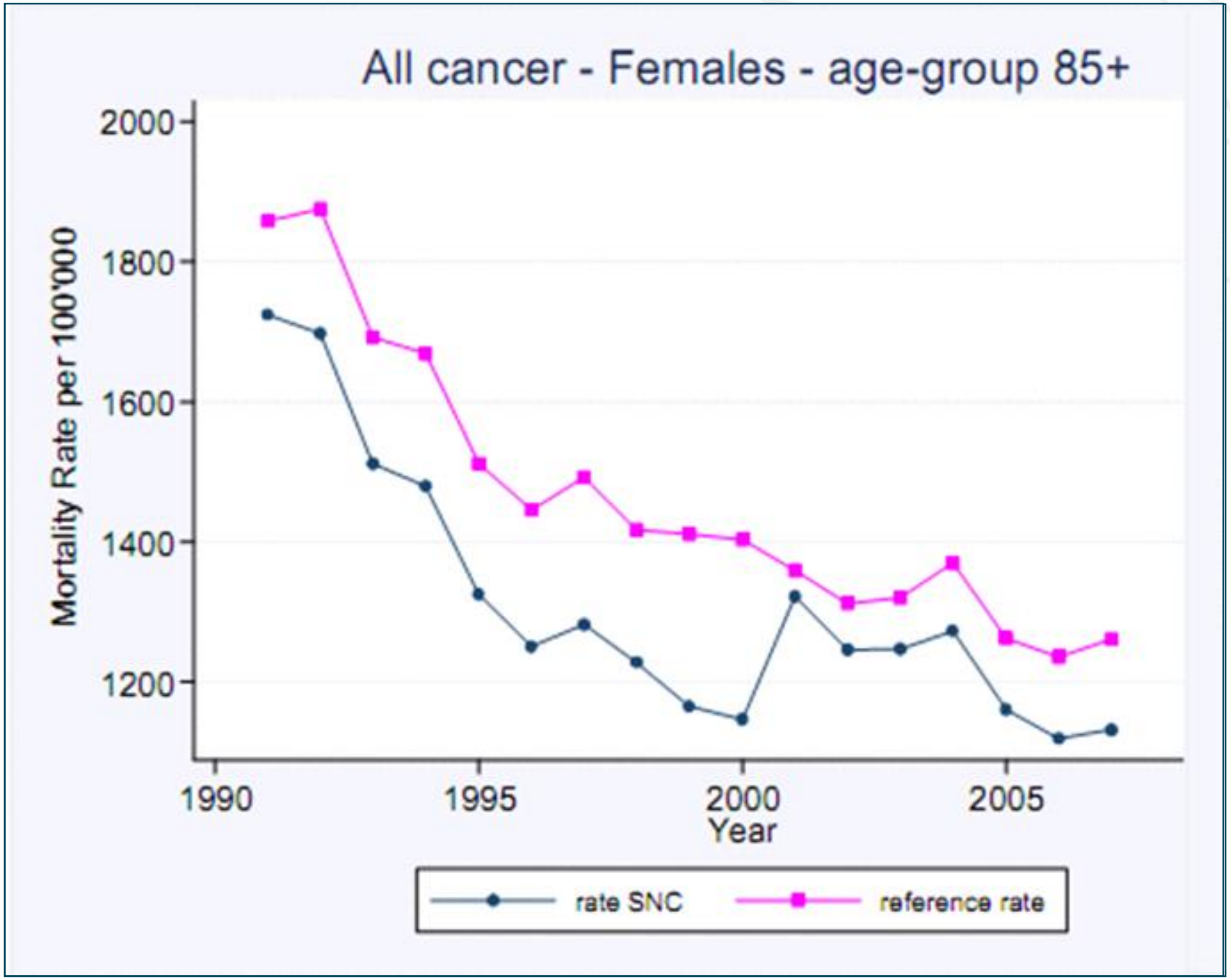
UCL Institute of Child Health

katie.harron.10@ucl.ac.uk

Background

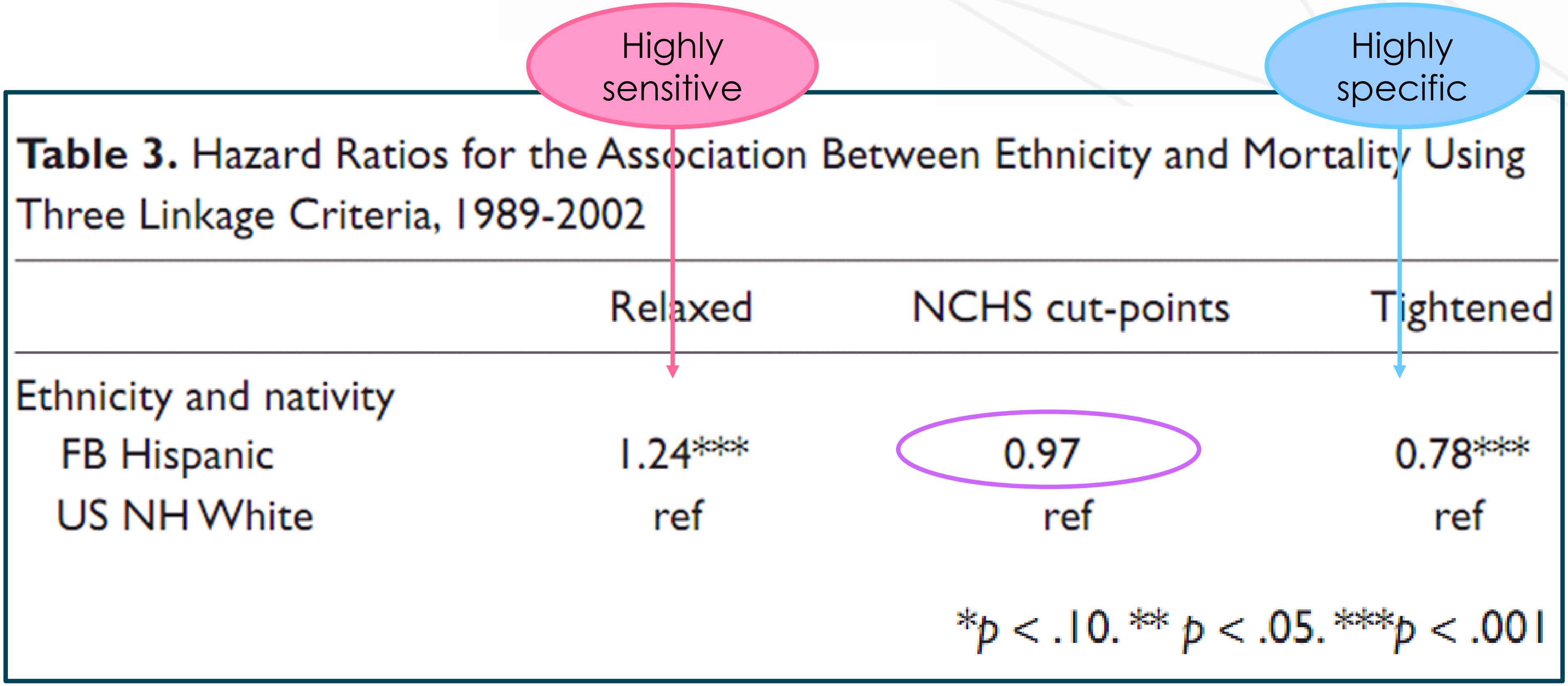
		Match status	
		Match (pair from same subject)	Non-match (pair from different subjects)
Link status	Link	Identified match	False match
	Non-link	Missed match	Identified non-match

Background



Schmidlin K et al (2013) Impact of unlinked deaths and coding changes on mortality trends in the Swiss National Cohort. BMC Med Inform Decis Mak 13 (1):1

Background



Highly sensitive

Highly specific

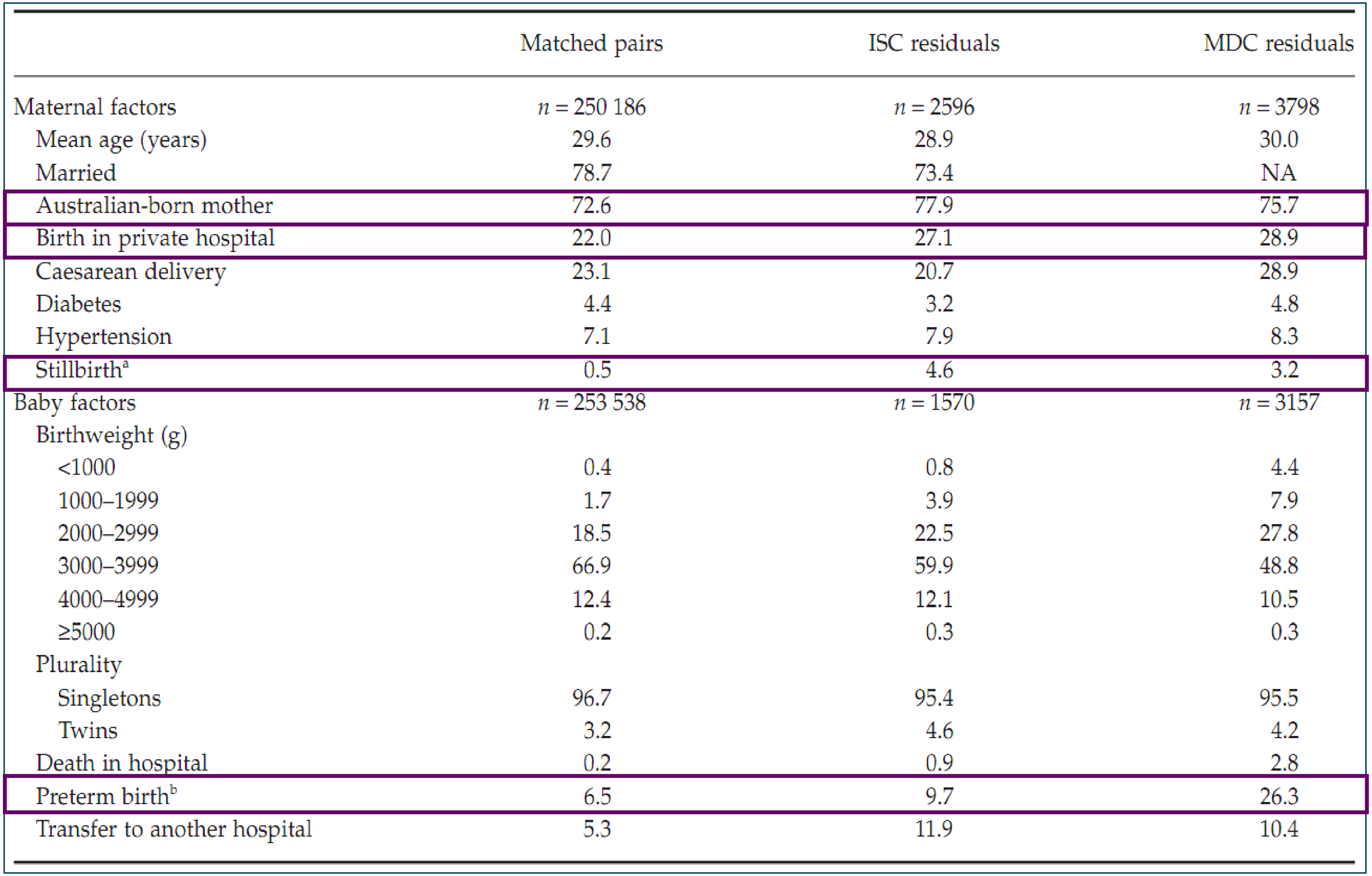
Table 3. Hazard Ratios for the Association Between Ethnicity and Mortality Using Three Linkage Criteria, 1989-2002

	Relaxed	NCHS cut-points	Tightened
Ethnicity and nativity			
FB Hispanic	1.24***	0.97	0.78***
US NH White	ref	ref	ref

*$p < .10$. ** $p < .05$. ***$p < .001$

Lariscy. Differential Record Linkage by Hispanic Ethnicity and Age in Linked Mortality Studies: Implications for the Epidemiologic Paradox *(2011, J Aging Health 2011)*

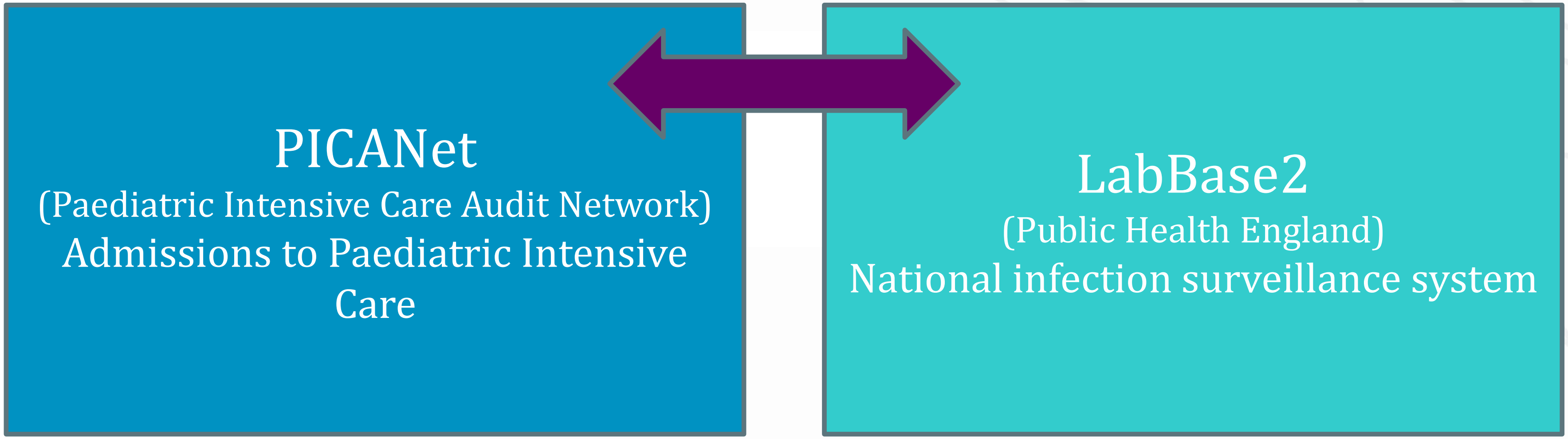
Background



	Matched pairs	ISC residuals	MDC residuals
Maternal factors	*n* = 250 186	*n* = 2596	*n* = 3798
Mean age (years)	29.6	28.9	30.0
Married	78.7	73.4	NA
Australian-born mother	72.6	77.9	75.7
Birth in private hospital	22.0	27.1	28.9
Caesarean delivery	23.1	20.7	28.9
Diabetes	4.4	3.2	4.8
Hypertension	7.1	7.9	8.3
Stillbirth[a]	0.5	4.6	3.2
Baby factors	*n* = 253 538	*n* = 1570	*n* = 3157
Birthweight (g)			
<1000	0.4	0.8	4.4
1000–1999	1.7	3.9	7.9
2000–2999	18.5	22.5	27.8
3000–3999	66.9	59.9	48.8
4000–4999	12.4	12.1	10.5
≥5000	0.2	0.3	0.3
Plurality			
Singletons	96.7	95.4	95.5
Twins	3.2	4.6	4.2
Death in hospital	0.2	0.9	2.8
Preterm birth[b]	6.5	9.7	26.3
Transfer to another hospital	5.3	11.9	10.4

Ford et al 2006. "Characteristics of unmatched maternal and baby records in linked birth records and hospital discharge data." Paediatric and Perinatal Epidemiology **20**(4): 329-337.

Methods



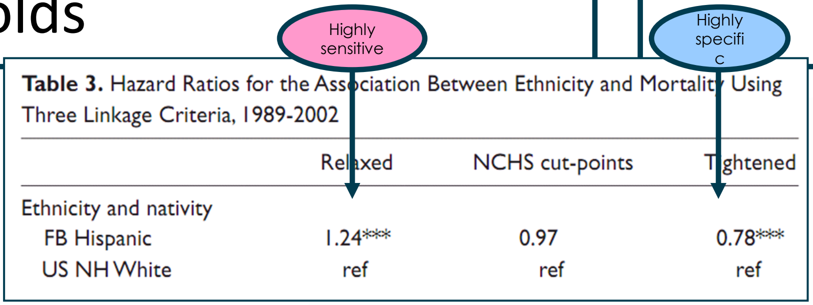
PICANet
(Paediatric Intensive Care Audit Network)
Admissions to Paediatric Intensive Care

LabBase2
(Public Health England)
National infection surveillance system

- □ Purpose: To estimate risk-adjusted infection rates
- □ Linkage using probabilistic match weights
- □ Four methods for evaluating linkage quality explored…

Methods

i) Sensitivity analysis using different probabilistic thresholds



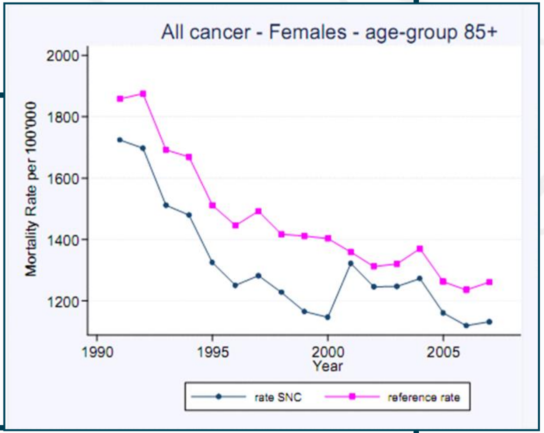
Highly sensitive

Highly specific

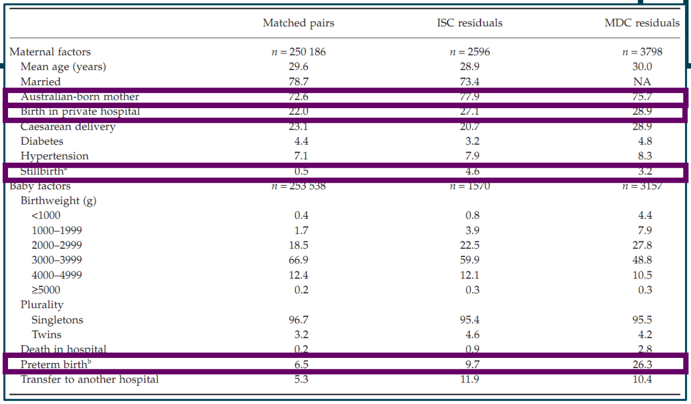
Table 3. Hazard Ratios for the Association Between Ethnicity and Mortality Using Three Linkage Criteria, 1989-2002

	Relaxed	NCHS cut-points	Tightened
Ethnicity and nativity			
FB Hispanic	1.24***	0.97	0.78***
US NH White	ref	ref	ref

ii) Subset of gold-standard data to quantify linkage bias



iii) Comparisons of linked and unlinked data

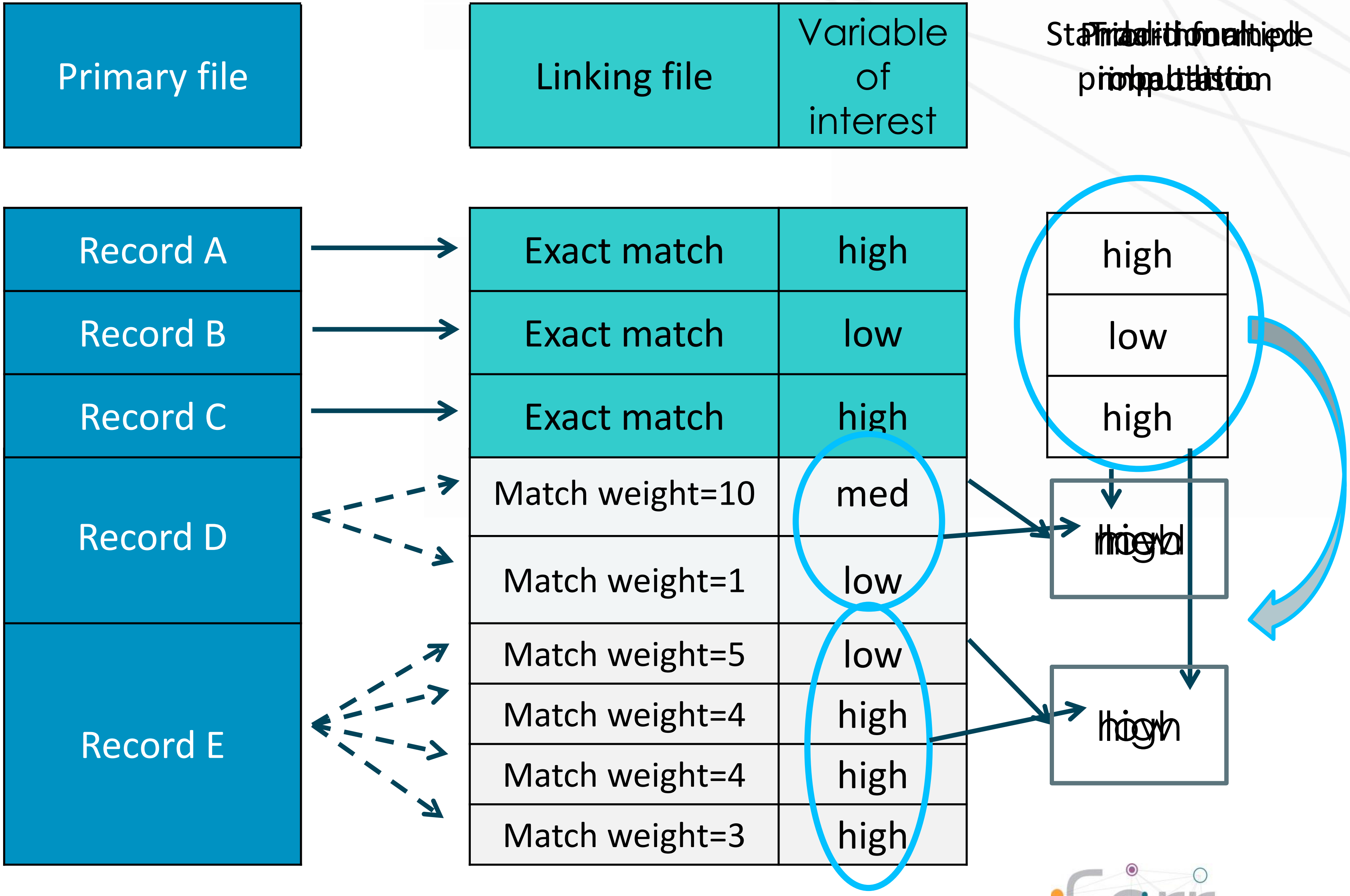


	Matched pairs	ISC residuals	MDC residuals
Maternal factors	*n* = 250 186	*n* = 2596	*n* = 3798
Mean age (years)	29.6	28.9	30.0
Married	78.7	73.4	NA
Australian-born mother	72.6	77.9	75.7
Birth in private hospital	22.0	27.1	28.9
Caesarean delivery	23.1	20.7	28.9
Diabetes	4.4	3.2	4.8
Hypertension	7.1	7.9	8.3
Stillbirth[a]	0.5	4.6	3.2
Baby factors	*n* = 253 538	*n* = 1570	*n* = 3157
Birthweight (g)			
<1000	0.4	0.8	4.4
1000–1999	1.7	3.9	7.9
2000–2999	18.5	22.5	27.8
3000–3999	66.9	59.9	48.8
4000–4999	12.4	12.1	10.5
≥5000	0.2	0.3	0.3
Plurality			
Singletons	96.7	95.4	95.5
Twins	3.2	4.6	4.2
Death in hospital	0.2	0.9	2.8
Preterm birth[b]	6.5	9.7	26.3
Transfer to another hospital	5.3	11.9	10.4

iv) Imputation for uncertain links

The Farr Institute of Health Informatics Research

Methods



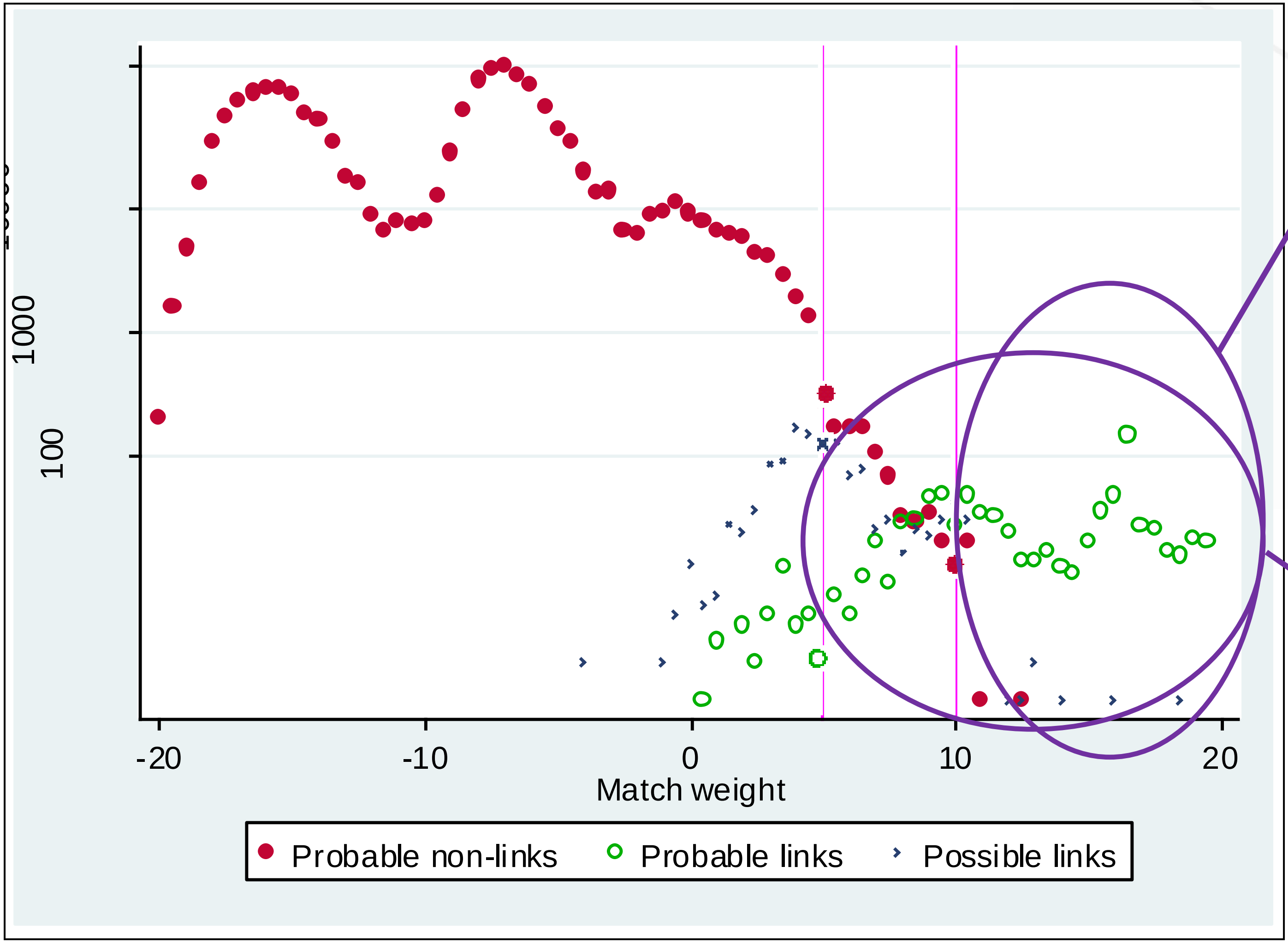
Primary file	Linking file	Variable of interest
Record A	Exact match	high
Record B	Exact match	low
Record C	Exact match	high
Record D	Match weight=10	med
	Match weight=1	low
Record E	Match weight=5	low
	Match weight=4	high
	Match weight=4	high
	Match weight=3	high

Standard multiple imputation
Traditional probabilistic
Prior-informed imputation
high
low
high
high
low
med
high
low

Goldstein H, Harron K, Wade A (2012) *The analysis of record-linked data using multiple imputation with data value priors*. Stat Med 31 (28):3481-3493

Results

i) Sensitivity analysis using different thresholds



Conservative threshold:

5249 admissions linked to 6083 specimens

Relaxed threshold:

7148 admissions linked to 8415 specimens

The Farr Institute of Health Informatics Research

Results

ii) Subset of gold-standard data

Complete, identified data from two laboratories

Threshold	Number of links identified	False matches	Missed-matches	Sensitivity	Positive Predictive Value
Gold-standard	426				
Relaxed (5)	492	75	9	417/426= 98%	417/492= 85%
Conservative (10)	418	26	34	392/426= 92%	392/418= 94%

3.8% infection rate; -1.9% bias

4.5% infection rate; 15.5% bias

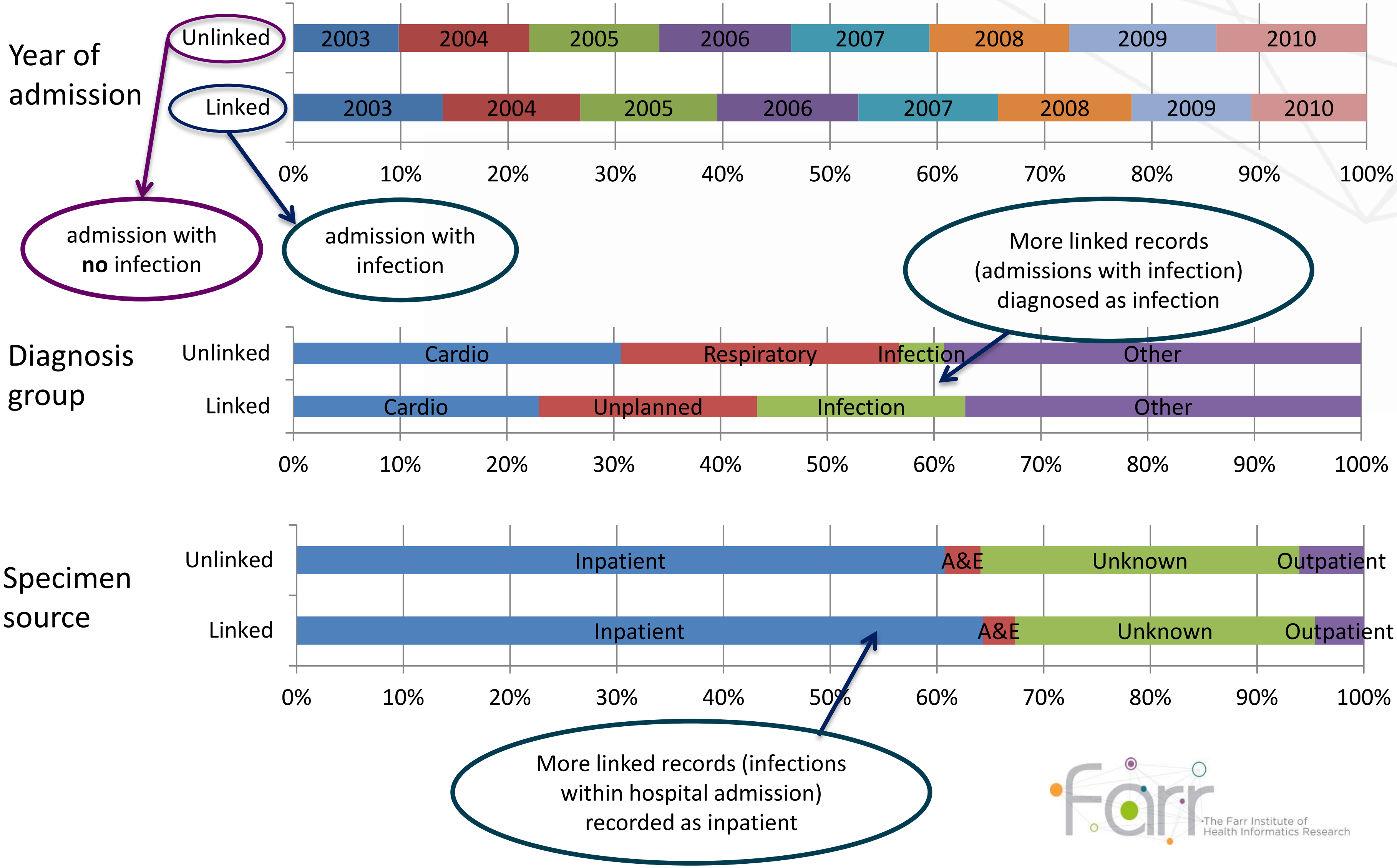
3.9% infection rate

The Farr Institute of Health Informatics Research

Results

iii) Comparison of linked and unlinked data characteristics

+ age, length of stay, sex



Unlinked → admission with **no** infection

Linked → admission with infection

More linked records (admissions with infection) diagnosed as infection

More linked records (infections within hospital admission) recorded as inpatient

The Farr Institute of Health Informatics Research

Results

iv) Imputation for uncertain links

Imputation: 424 links; 3.9% infection rate

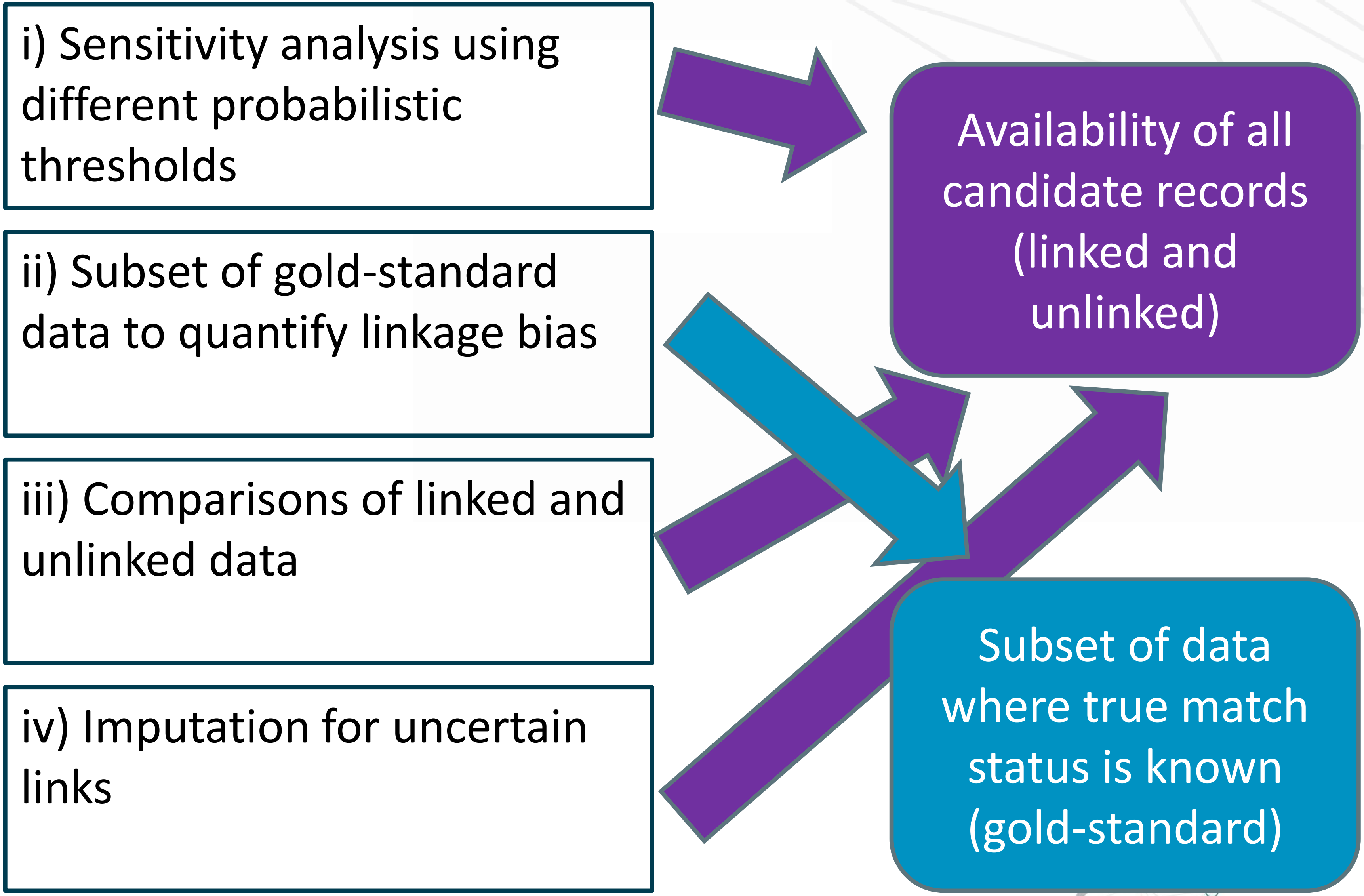
Gold-standard: 426 links; 3.9% infection rate

Probabilistic conservative: 418 links; 3.8% infection rate

Probabilistic relaxed: 492 links; 4.5% infection rate

The Farr Institute of Health Informatics Research

Conclusions



i) Sensitivity analysis using different probabilistic thresholds

ii) Subset of gold-standard data to quantify linkage bias

iii) Comparisons of linked and unlinked data

iv) Imputation for uncertain links

Availability of all candidate records (linked and unlinked)

Subset of data where true match status is known (gold-standard)

Conclusions

❑Evaluation of linkage quality

→ vital

❑Existing situation

→ physical separation of identifiers and clinical data

❑What is needed

→ all candidate records (including unlinked data) transferred alongside non-disclosive data

The Farr Institute of Health Informatics Research

Acknowledgements

Co-investigators: Ruth Gilbert, Berit Muller-Pebody, Harvey Goldstein, Angie Wade

John Hartley, Paul Lock and Joanna Bell - Great Ormond Street Hospital

Michael Millar and David Ball - Royal London Hospital

Jim Gray - Birmingham Children's Hospital

PICANet team, Roger Parslow and Tom Fleming

Quen Mok, Tracy Ball and the CATCH team

Funding from the NIHR Health Technology Assessment Programme (ref 08/13/47 http://www.hta.ac.uk/1867) *The views and opinions expressed therein are those of the authors and do not necessarily reflect those of the HTA programme, NIHR, NHS or the Department of Health.*

Contact: katie.harron.10@ucl.ac.uk

UCL

The Farr Institute of Health Informatics Research