LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

Enhancing observational data through linkage: opportunities and challenges

Dr Katie Harron, Sir Henry Wellcome Postdoctoral Fellow

London School of Hygiene and Tropical Medicine

Administrative Data Research Centre England

Farr The Farr Institute London

wellcome trust Fellow

Data linkage

Each person in the world creates a Book of Life.

This Book starts with birth and ends with death.

Its pages are made up of the records of the principal events in life.

Record linkage is the name given to the process of assembling the pages of this Book, into a volume.

Dunn, 1946

Opportunities and challenges using administrative / clinical / routine data

+ population-level resource
+ detailed longitudinal healthcare trajectories
+ allows evaluation of rare events / specific subgroups
+ potentially lower risk of selection bias
+ answer novel research questions

Answering novel research questions

Deep vein thrombosis and air travel: record linkage study

C W Kelman, M A Kortt, N G Becker, Z Li, J D Mathews, C S Guest, C D J Holman

Abstract

Objective To investigate the time relations between lon

De

ap

Set

Par

ver

arr

Re

increased for only two weeks after a long haul flight; 46 Australian citizens and 200 non-Australian citizens had an episode of venous thromboembolism during this so called hazard period. The relative risk during this period for Australian citizens was 4.17 (95% confidence interval, 2.94 to 5.40), with 76% of cases (n = 35) attributable to the preceding flight. A "healthy traveller" effect was observed, particularly for Australian citizens.

Conclusions The annual risk of venous thromboembolism is increased by 12% if one long haul flight is taken yearly. The average risk of death from flight related venous thromboembolism is small compared with that from motor vehicle crashes and injuries at work. The individual risk of death from flight related venous thromboembolism for people with certain pre-existing medical conditions is, however, likely to be greater than the average risk of 1 per 2 million for passengers arriving from a flight. Airlines and health authorities should continue to advise passengers on how to minimise risk.

Conclusions The annual risk of venous thromboembolism is increased by 12% if one long haul flight is taken yearly.

pulmonary embolism after long flights has brought the issue to public attention.

The incidence of venous thromboembolism ranges

10-30% of patients with venous thromboembolism.[9]

International air travel has increased to around 1.56 billion person trips each year.[10] At any one time an estimated 4000 Australians are on international flights, and more than 30 000 make short domestic flights each day.

Since 1970, Australia has kept electronic data on arrivals and departures of international travellers. The state of Western Australia uses record linkage under well developed protocols to protect patient privacy.[11] Most Western Australian residents live in Perth, and flight times from there to other major airports are long. We investigated the relation between international air travel and venous thromboembolism by linking Western Australian hospital data with records on air travel.

Participants and methods

Data included coded personal identifiers, age, sex, arrival and departure dates, and nationality of the trav-

Commonwealth Department of Health and Ageing,

National Centre for Epidemiology and Population Health, Australian National University, Canberra, ACT 0200

N G Becker *professor of biostatistics*

Z Li *postdoctoral fellow*

C S Guest *visiting fellow*

School of Population Health, University of Western Australia, Perth, WA 6009, Australia

C D J Holman *chair in public health*

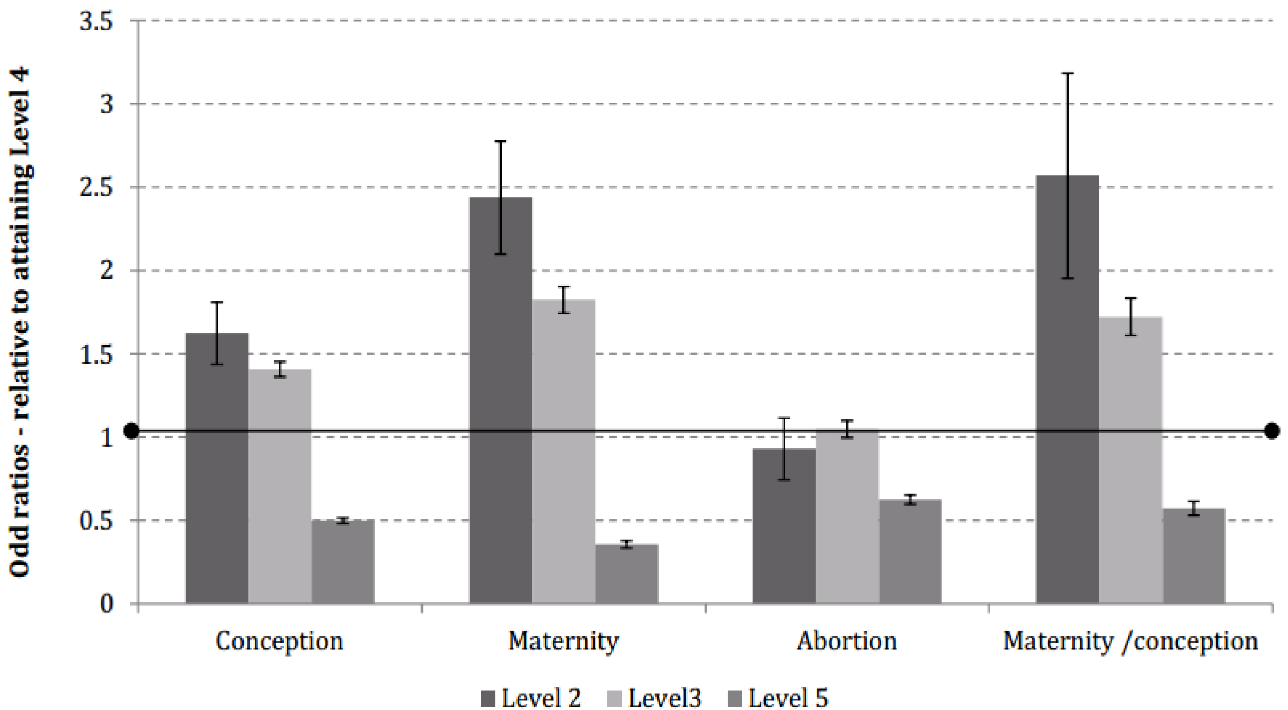
Correspondence to: C W Kelman christopher.kelman@health.gov.au

Electronic data on flight arrivals and departures

Hospitalisations data

Answering novel research questions

Figure 3.3 Raw differences in conception behaviour by the end of Year 11, by Key Stage 2 English scores: estimated odds ratios relative to attaining Level 4



CAYT Impact Study: Report No. 6

Claire Crawford
Jonathan Cribb
Elaine Kelly

National Pupil Database (NPD)

Office for National Statistics (ONS) conceptions data

Opportunities and challenges using administrative / clinical / routine data

+ population-level resource
+ detailed longitudinal healthcare trajectories
+ allows evaluation of rare events / specific subgroups
+ potentially lower risk of selection bias
+ answer novel research questions

- uncertainty about data quality
- different ways to code the same outcome
- information found across a number of fields
- sometimes lack of consistency

Linkage to overcome data quality issues

+ Allows triangulation of outcomes
+ Improves ascertainment

→ Example 1:
Mother-baby linkage in English hospital data

- Complete, accurate identifiers not always available
- Potential for introducing bias due to linkage error

→ Example 2:
Evaluating error in linkage of intensive care / laboratory data

Mother-baby linkage in NHS

- Mother and baby records not routinely linked in data within the English National Health Service (NHS)

- Linked maternal-baby data is available in other countries
  - e.g. Scotland, Canada, Australia, Netherlands, US

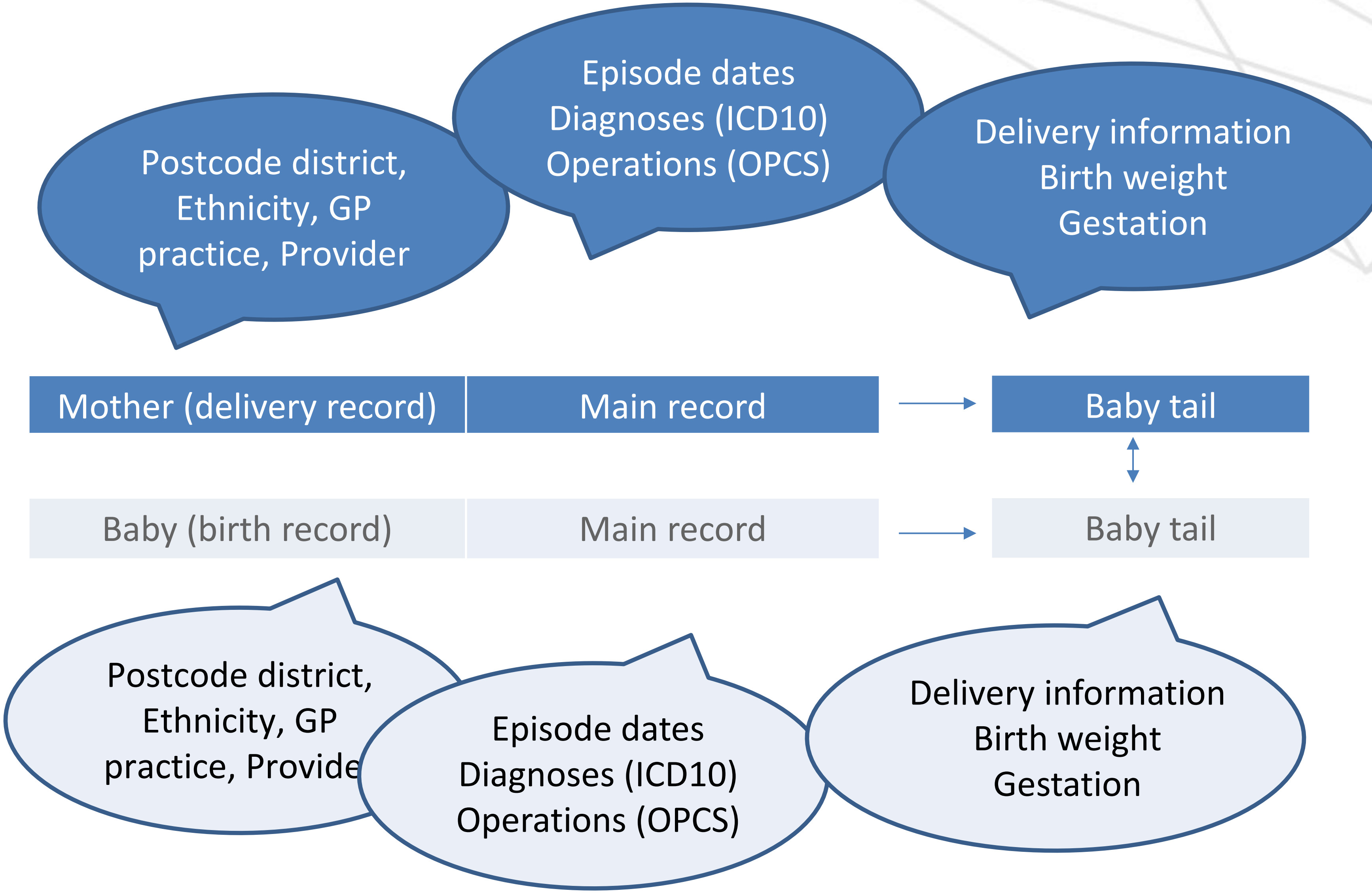
- Linkage of prospective data planned for future in England

We evaluated whether linkage of retrospective data from maternal and baby records could be used to address data quality issues in English hospital data

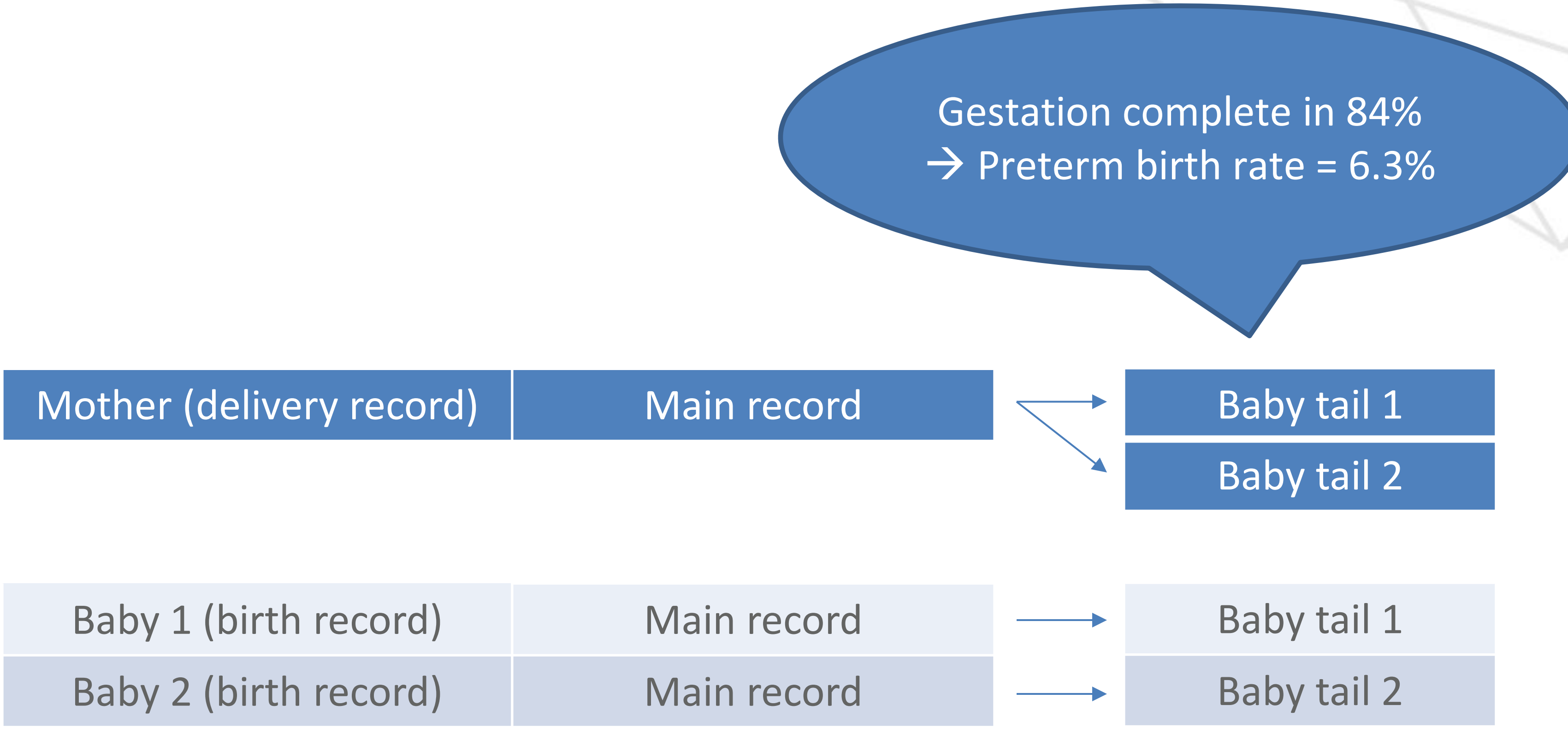
- Deterministic and probabilistic linkage of “indirect” identifiers

- Non-disclosive variables contained in both maternal and baby records

Example: mother-baby linkage



Example: mother-baby linkage



Can linkage with baby records help improve ascertainment?

Example: mother-baby linkage

ICD10: Z371 single still birth
Z373 twins, one live on still
Z374 twins, both stillborn
Z377 other multiple, stillborn
O364 maternal care for intrauterine death

0.55%

0.49%

Birth status: (live or still)

Mother (delivery record) | Main record → Baby tail 1, Baby tail 2

Baby 1 (birth record) | Main record → Baby tail 1

Baby 2 (birth record) | Main record → Baby tail 2

Example: mother-baby linkage

ICD10: Z371 single still birth
Z373 twins, one live on still
Z374 twins, both stillborn
Z377 other multiple, stillborn
O364 maternal care for intrauterine death

0.55%

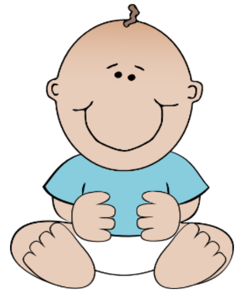
0.49%

Birth status: (live or still)

		ICD		
		Live	**Still**	
Birth status	**Live**	99.34%	0.17%	*668,141*
	Still	0.12%	0.38%	*3295*
		667,797	*3639*	*675,734*

Can linkage with baby records help resolve inconsistencies?

Linkage



Baby records 2012
N = 673,055

Maternal records 2012
N=671,436

Deterministic linkage:
GP practice
Maternal age
Birth weight
Gestation
Birth order
Sex of baby

280,939 linked baby records (42%)

Linkage

391,705 remaining unlinked baby records

Probabilistic linkage

380,164 link

GP pra
Matern
Birth weight
Gestation
Birth order

Clinical variables
First antenatal assessment date
Estimated delivery date
Gestation at first antenatal assessment
Delivery place (actual)
Delivery place (intended)
livery
delivery
hod
Birth
umber of babie
Episode start date
Episode end date

Total combining deterministic and probabilistic:
[illegible]0,401 linked baby-mother records (98% of babies)

Partial Identifiers
Postcode district
Ethnicity

Combining information from baby and mother records

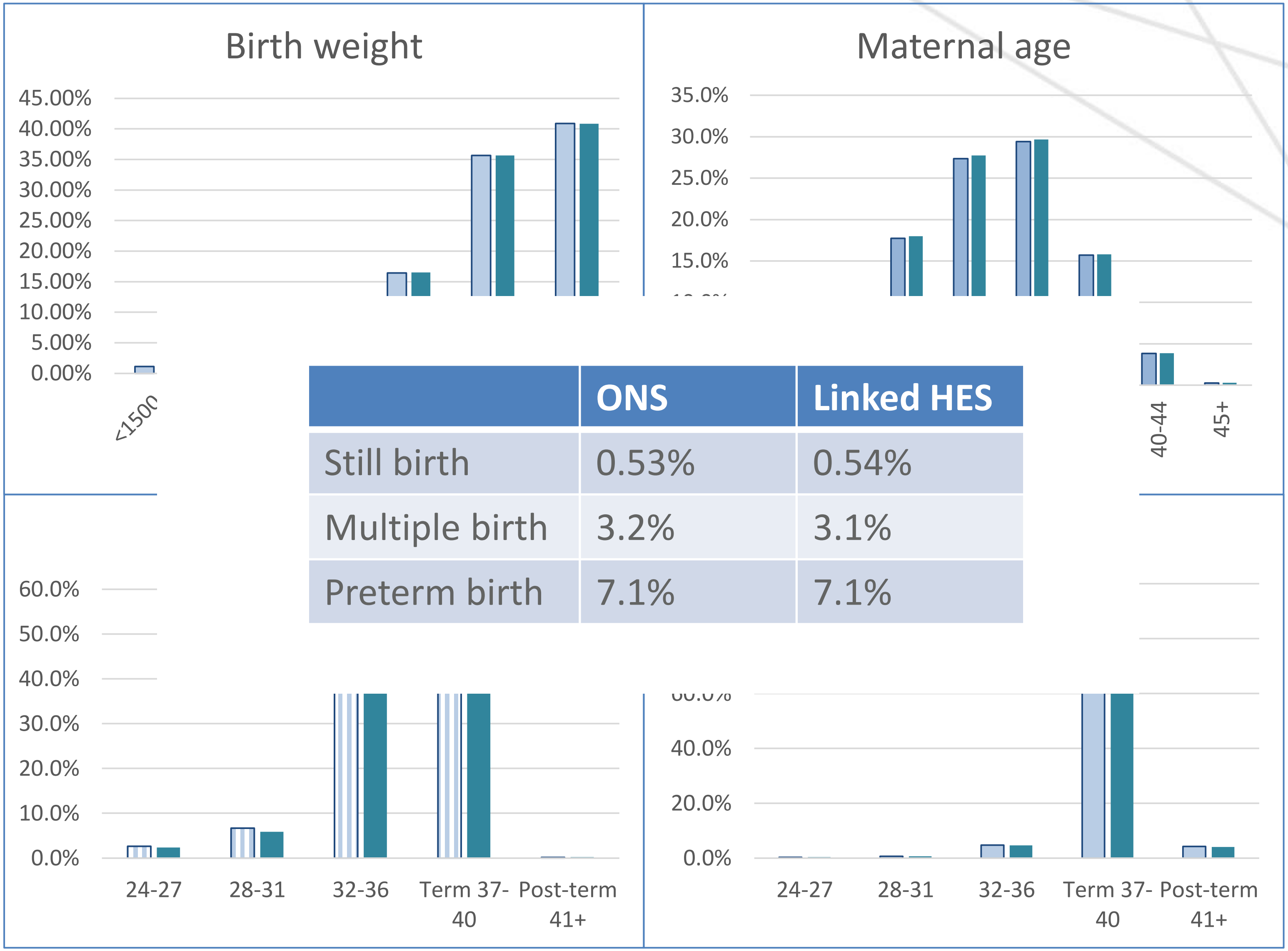
Ascertainment

- Completeness of gestation increases from 84% → 92%
- Preterm birth rate increases from 6.1% → 6.7%
  - Further increases to 6.9% using ICD10 code for preterm birth O60 in baby record

Inconsistencies

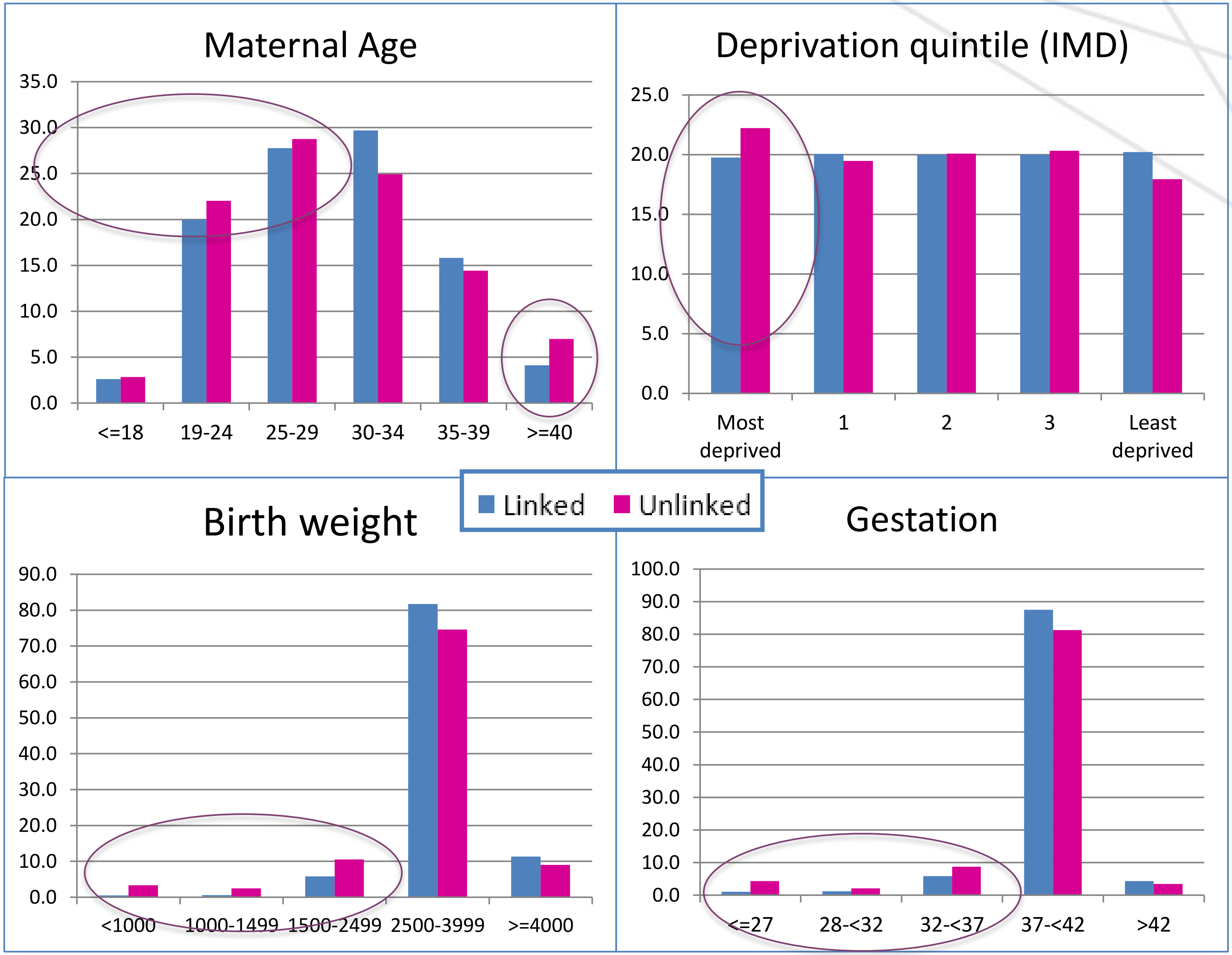
- 800/1558 stillbirth conflicts resolved through information held on baby record
  - Checking ICD10 codes, birth status, length of stay
  - 0.1% of records unresolved

Checking external validity



	ONS	Linked HES
Still birth	0.53%	0.54%
Multiple birth	3.2%	3.1%
Preterm birth	7.1%	7.1%

Comparing linked (98%) vs unlinked (2%)



Linkage error

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

The linkage problem

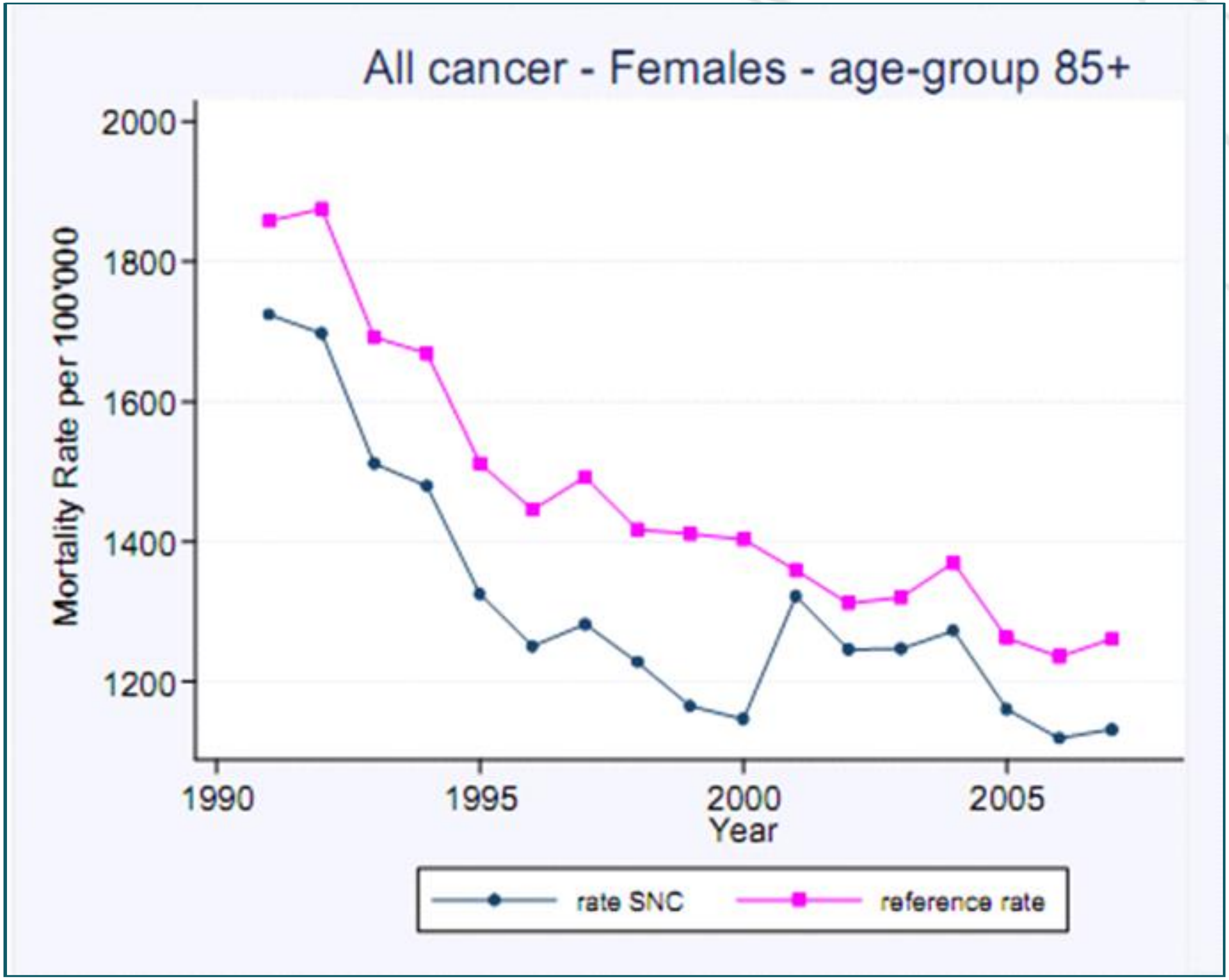
- Small amounts of linkage error can result in substantially biased results

- False matches

  → introduce variability and weaken the association between variables – bias to the null

- Missed matches

  → reduce our sample size and result in a loss of power – potential selection bias



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

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)*

	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

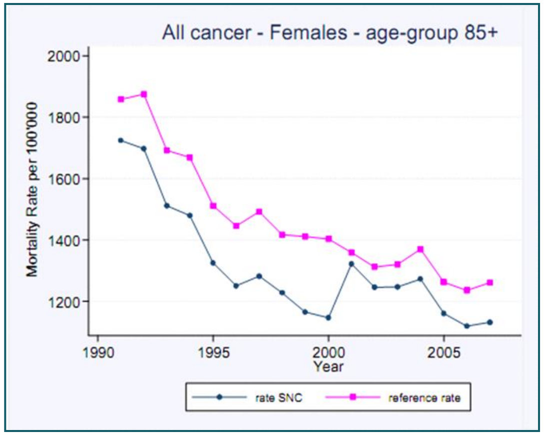
Ford JB, Roberts CL, Taylor LK (2006) Characteristics of unmatched maternal and baby records in linked birth records and hospital discharge data. Paediatr Perinat Ep 20 (4):329-337

Differential linkage – why?

- Data quality differs by patient group:
  - Bohensky et al 2010. Data Linkage: A powerful research tool with potential problems. *BMC Health Services Research*
- Unknown/estimated dates of birth
  - Unconscious, frail, dementia,
- Unconventional surnames
- Address issues
  - Communal establishments
  - Visitor / tourist / traveller
- Misleading information
  - Drug user, parent withholding details
- Multiple births

Evaluating bias due to linkage error

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



ii) 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

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) Statistical / missing data methods - imputation for uncertain links

Example: Bloodstream infection in paediatric intensive care units

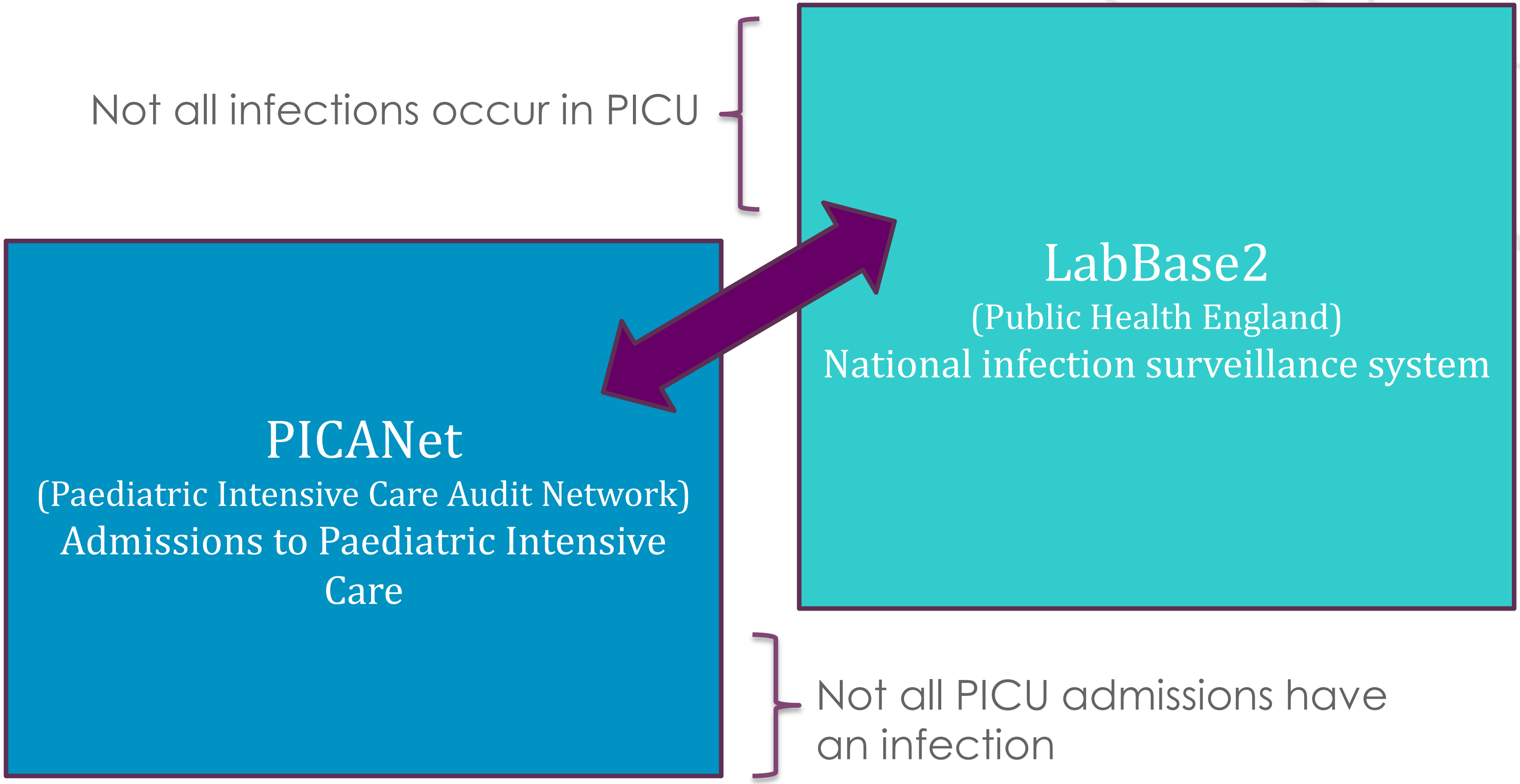
PICANet
(Paediatric Intensive Care Audit Network)
Admissions to Paediatric Intensive Care

LabBase2
(Public Health England)
National infection surveillance system

AIM: To evaluate trends in risk-adjusted infection rates in PICU

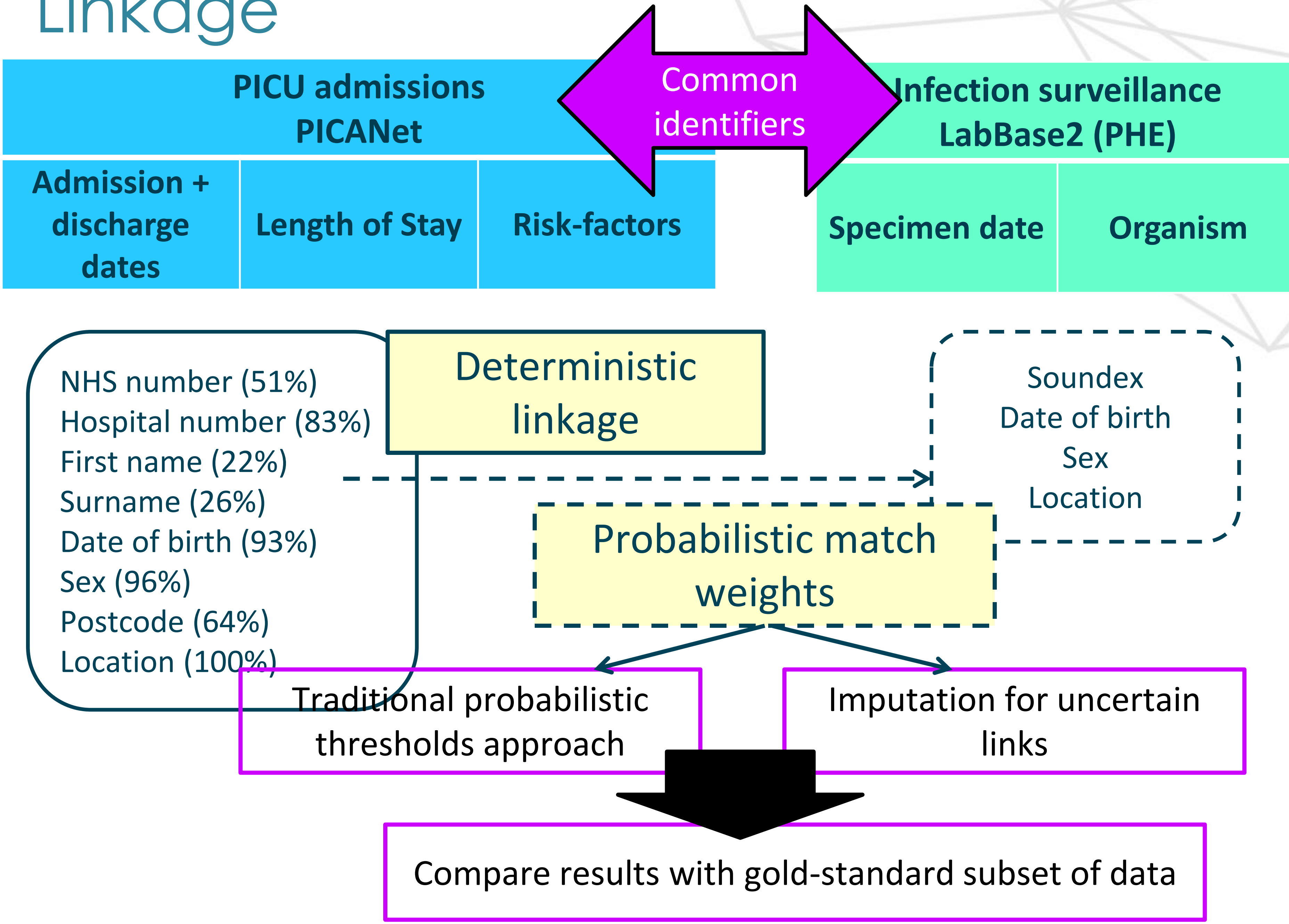
- Linkage using deterministic and probabilistic linkage
- Explored bias due to linkage error using
  - gold-standard data
  - sensitivity analyses
  - imputation

Example: Bloodstream infection in paediatric intensive care

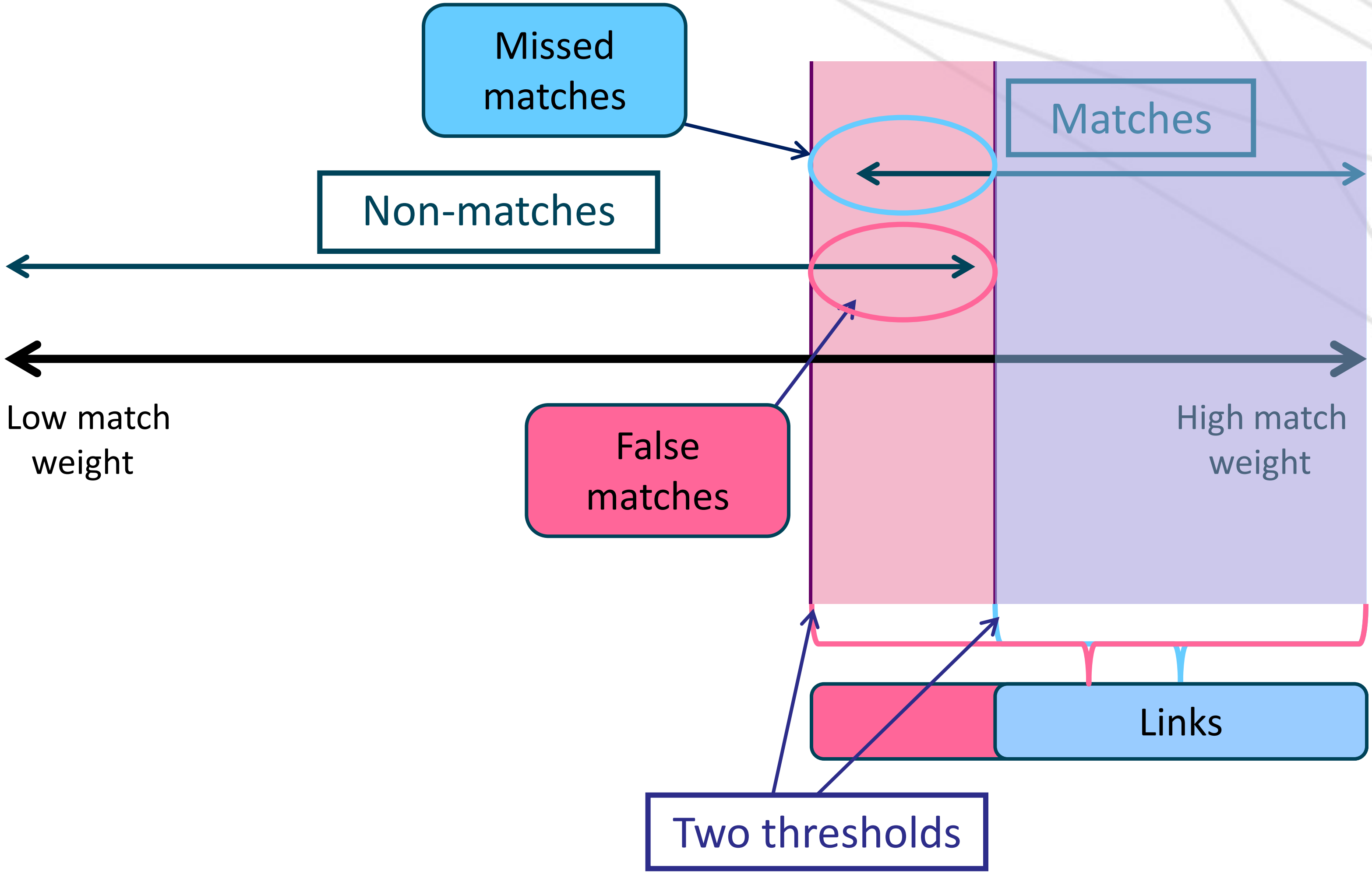


Comparing characteristics of linked and unlinked data not helpful in this context

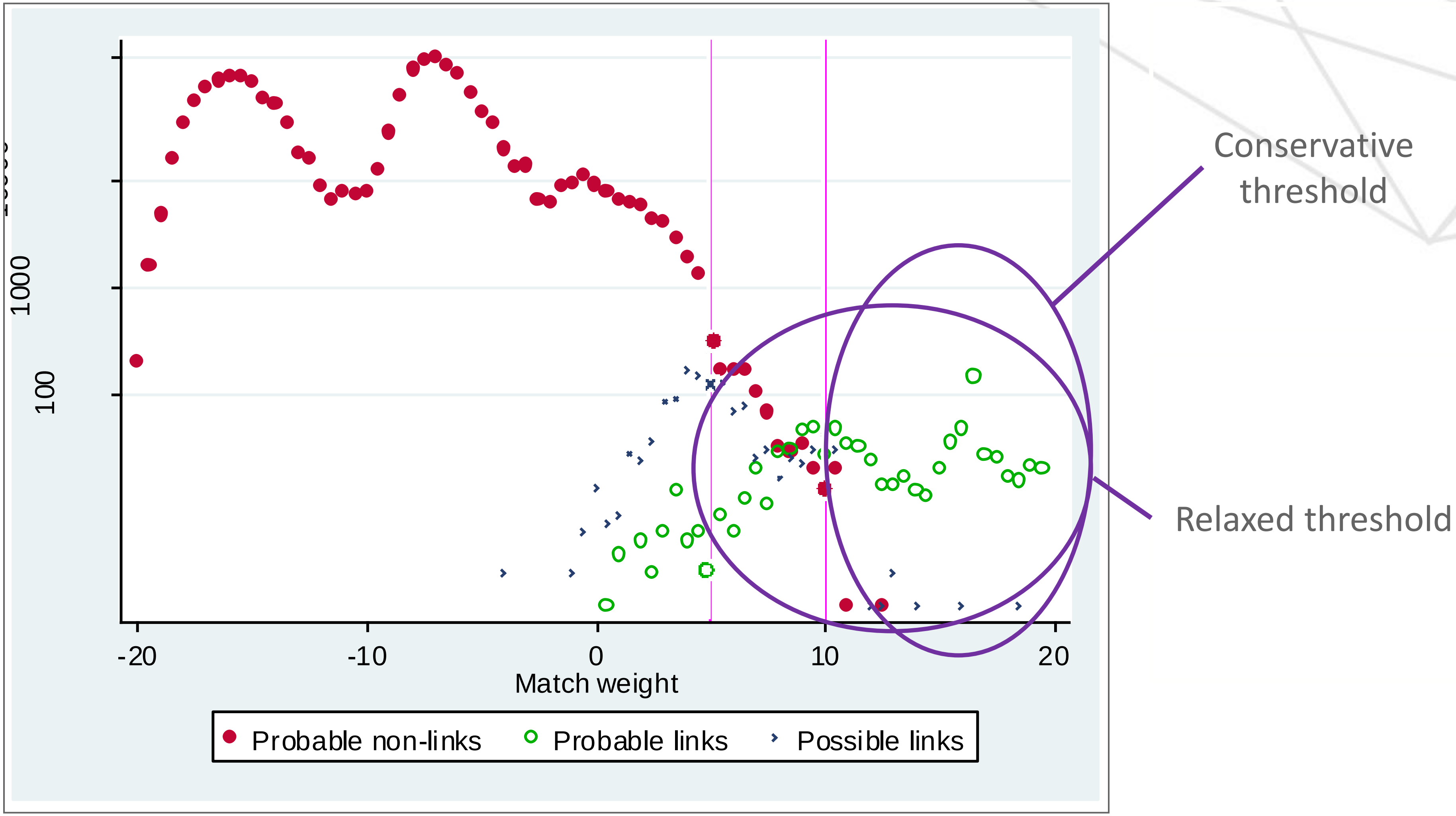
Linkage


PICU admissions
PICANet
Common identifiers
Infection surveillance
LabBase2 (PHE)
Admission + discharge dates
Length of Stay
Risk-factors
Specimen date
Organism
NHS number (51%)
Hospital number (83%)
First name (22%)
Surname (26%)
Date of birth (93%)
Sex (96%)
Postcode (64%)
Location (100%)
Deterministic linkage
Soundex
Date of birth
Sex
Location
Probabilistic match weights
Traditional probabilistic thresholds approach
Imputation for uncertain links
Compare results with gold-standard subset of data

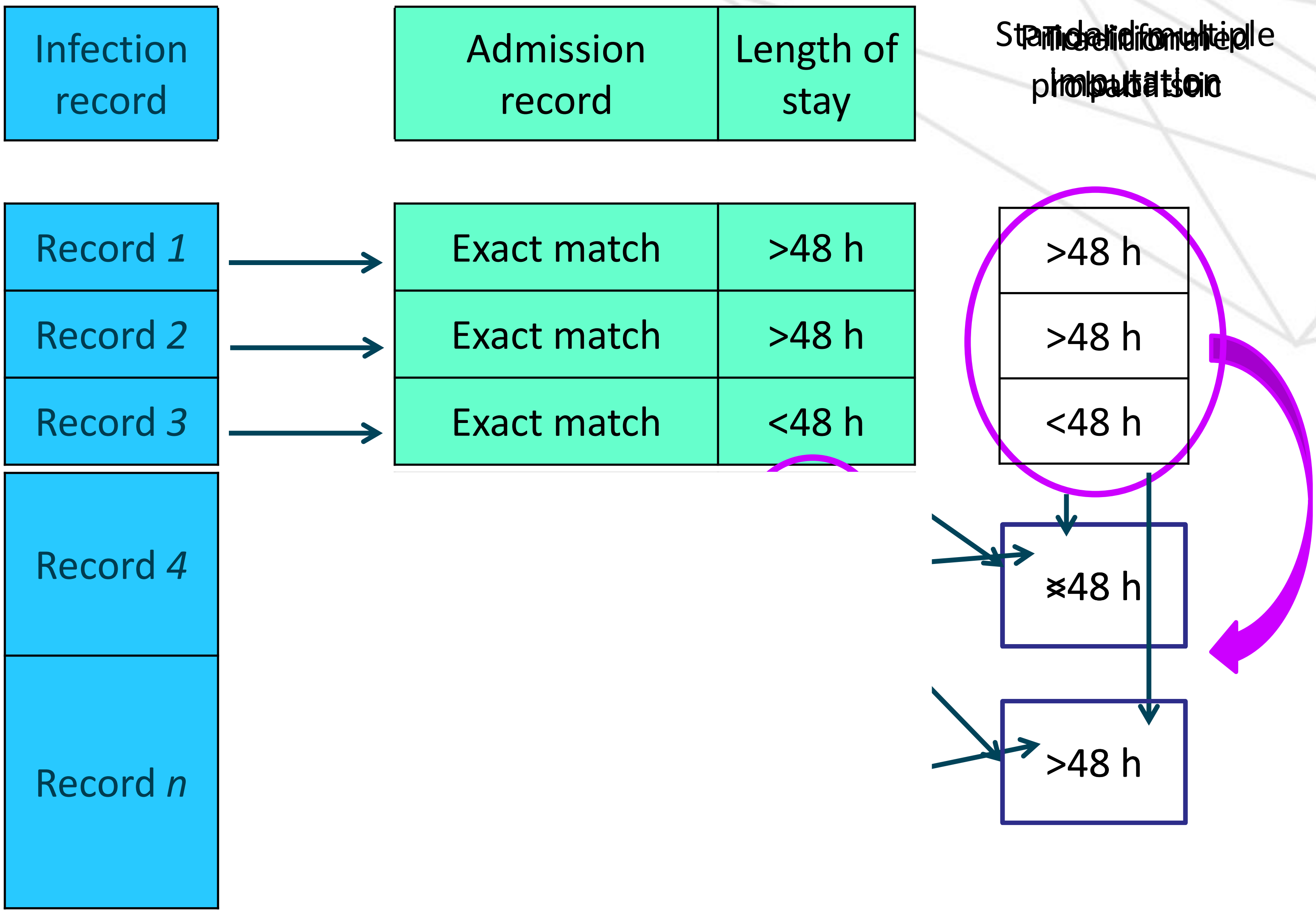
Choosing a threshold weight



Evaluation: sensitivity analysis



Imputation for uncertain links

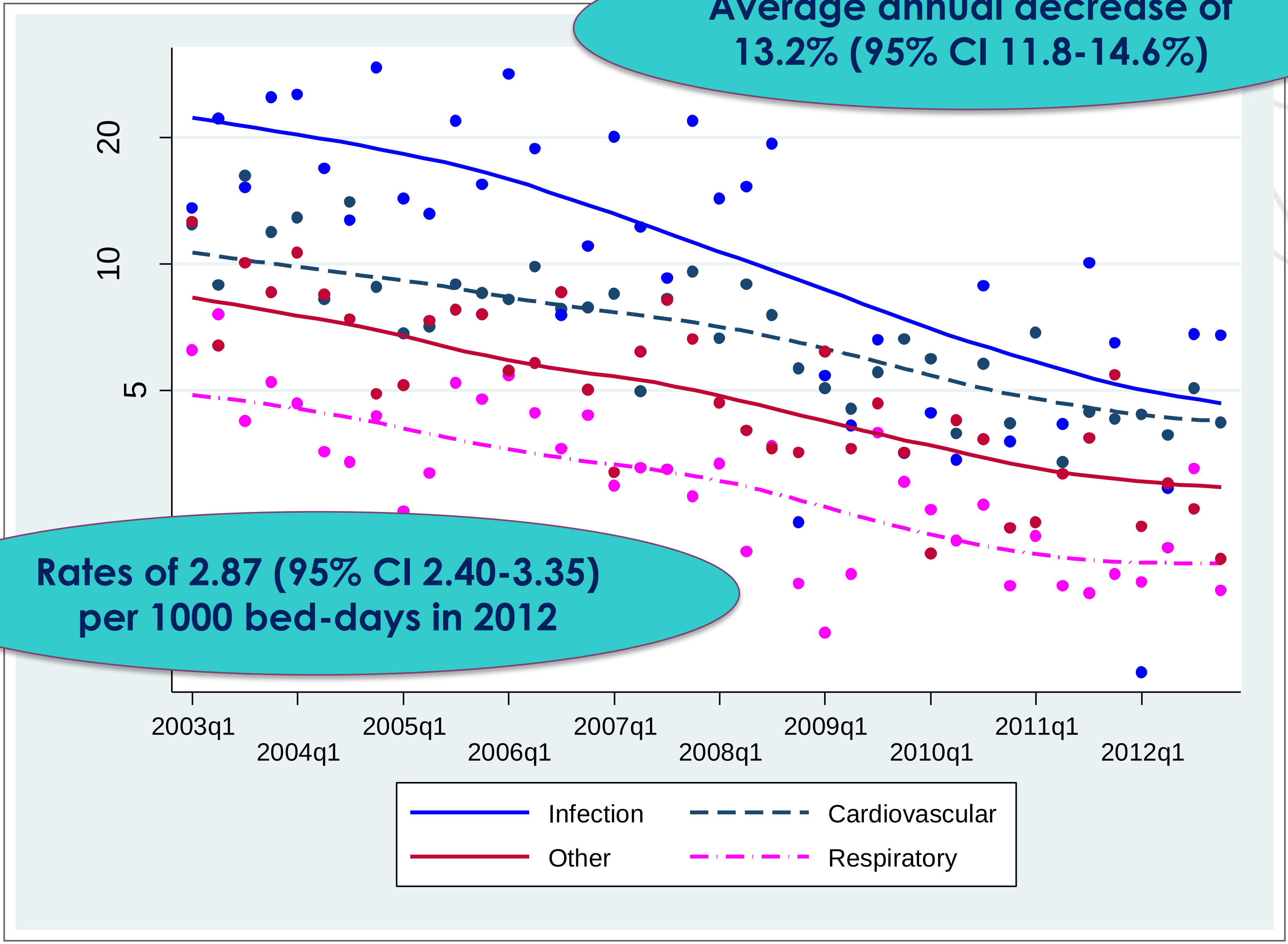


Goldstein et al *Stat Med* 2012;31(28):3481-3493
Harron et al *BMC Med Res Method* 2014;14(1):36

Evaluation: gold-standard data

Threshold	Number of links identified	Infection rate (%)	Bias (%)
Gold-standard	426	3.9	
Relaxed (5)	492	4.5	15.5
Conservative (10)	418	3.8	-1.9
Imputation	424	3.9	-0.5

Results



Average annual decrease of 13.2% (95% CI 11.8-14.6%)

Rates of 2.87 (95% CI 2.40-3.35) per 1000 bed-days in 2012

Reporting studies using linked data

- Readers should also understand
  - Any limitations of the data
    - e.g. still birth
  - Processes by which errors occur
    - e.g. ethnic group
  - Implications for analyses
    - e.g. potential selection bias

- Transparency and clear reporting helps interpretation
  - Can be a challenge to obtain information on linkage process

Reporting

REporting of studies Conducted using Observational Routinely-collected Data

- RECORD initiative: reporting guidelines for studies conducted using routinely-collected health data
  - record-statement.org
- Addresses unique challenges associated with using data collected primarily for reasons other than research
- Specific section on data linkage

PLOS | MEDICINE   Browse   Publish   About

The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) Statement

Eric I. Benchimol, Liam Smeeth, Astrid Guttmann, Katie Harron, David Moher, Irene Petersen, Henrik T. Sørensen, Erik von Elm, Sinéad M. Langan, RECORD Working Committee

Published: October 6, 2015 • DOI: 10.1371/journal.pmed.1001885

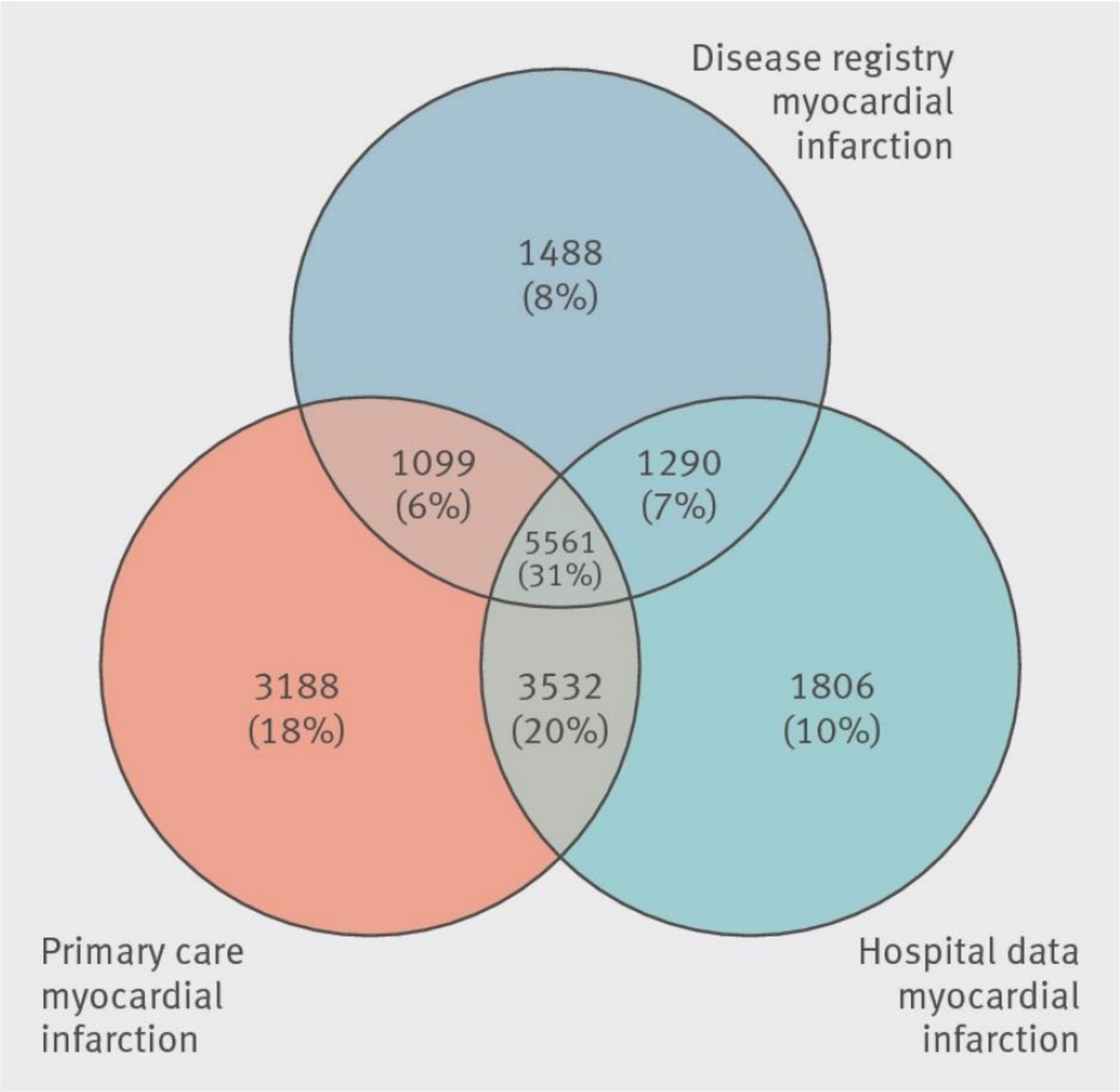
Reporting

REporting of studies Conducted using Observational Routinely-collected Data

RECORD 6.3: If the study involved linkage of databases, consider use of a flow diagram or other graphical display to demonstrate the data linkage process, including the number of individuals with linked data at each stage.

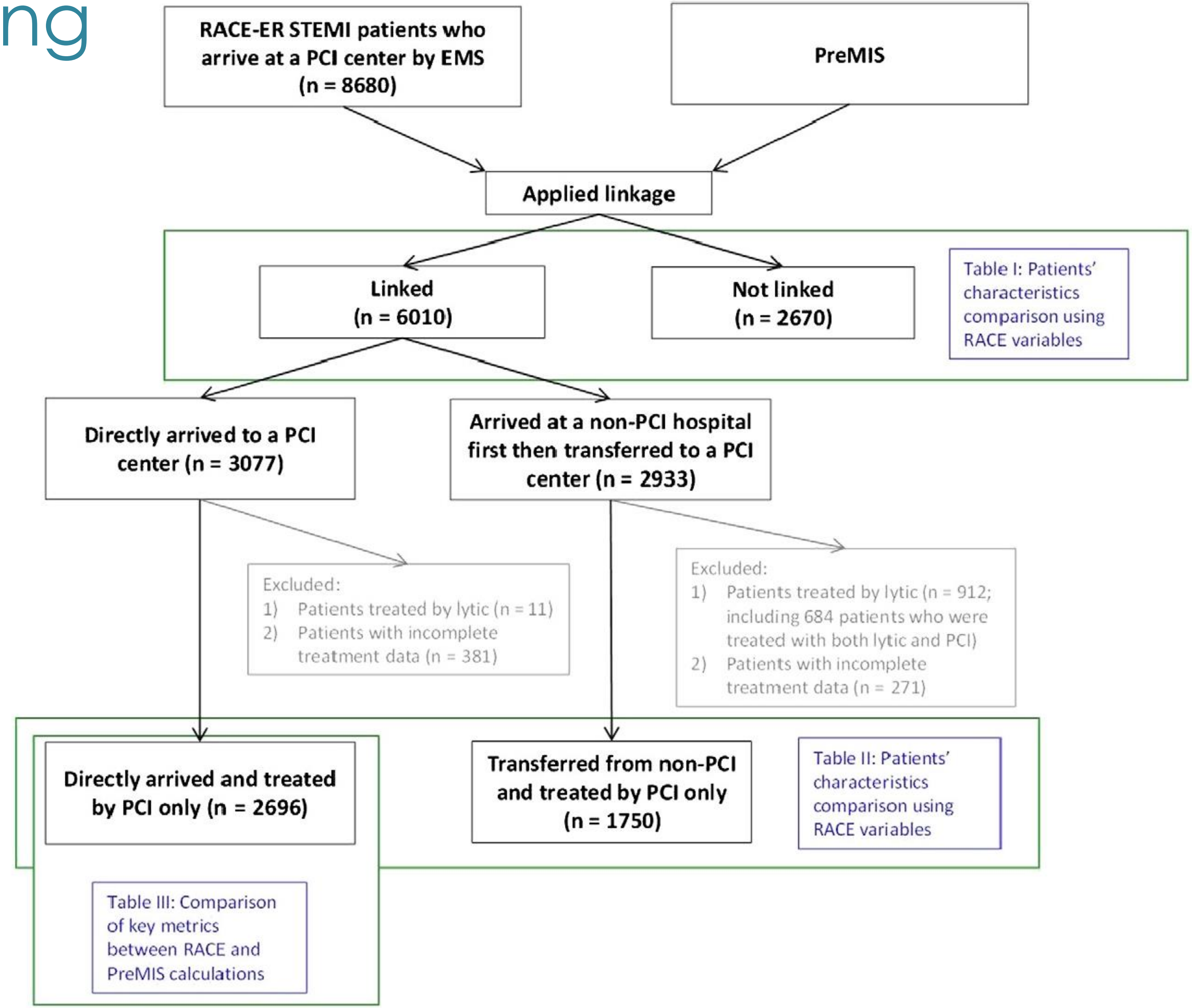
RECORD 12.3: State whether the study included person-level, institutional-level, or other data linkage across two or more databases. The methods of linkage and methods of linkage quality evaluation should be provided.

Reporting



Disease registry myocardial infarction
1488 (8%)
1099 (6%)
1290 (7%)
5561 (31%)
3188 (18%)
3532 (20%)
1806 (10%)
Primary care myocardial infarction
Hospital data myocardial infarction

Reporting



RACE-ER STEMI patients who arrive at a PCI center by EMS (n = 8680)
PreMIS
Applied linkage
Linked (n = 6010)
Not linked (n = 2670)
Table I: Patients' characteristics comparison using RACE variables
Directly arrived to a PCI center (n = 3077)
Arrived at a non-PCI hospital first then transferred to a PCI center (n = 2933)
Excluded:
1) Patients treated by lytic (n = 11)
2) Patients with incomplete treatment data (n = 381)
Excluded:
1) Patients treated by lytic (n = 912; including 684 patients who were treated with both lytic and PCI)
2) Patients with incomplete treatment data (n = 271)
Directly arrived and treated by PCI only (n = 2696)
Transferred from non-PCI and treated by PCI only (n = 1750)
Table II: Patients' characteristics comparison using RACE variables
Table III: Comparison of key metrics between RACE and PreMIS calculations

Summary

- Linkage can help to address data quality issues
  - Improve ascertainment of key risk-factors and outcomes
  - Triangulate outcomes and resolve inconsistencies
  - Highlights limitations in the data
- Understanding bias due to linkage error is important
  - Several approaches available for evaluating potential impact on results
  - Requires information on linkage process and unlinked records
  - Transparent reporting can aid interpretation
- Unfulfilled opportunities – linkage between health and other sectors on new scale
  - e.g. Brazil's 100 million cohort: socio-economic and health data

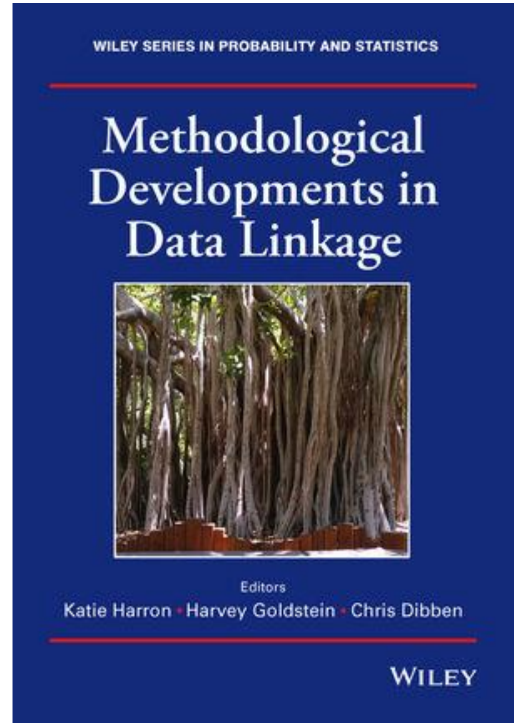
Farr The Farr Institute of Health Informatics Research

Acknowledgements and funding

Fellowship steering committee:

Jan van der Meulen,
Ruth Gilbert
David Cromwell
Astrid Guttmann
Harvey Goldstein

Thanks also to Hannah Knight and Ipek Gurol (Royal College of Obstetricians and Gynecologists)



This work was supported by funding from the Wellcome Trust (103975/Z/14/Z)

Hospital Episode Statistics were made available by the NHS Health and Social Care Information Centre (Copyright © 2012, Re-used with the permission of The Health and Social Care Information Centre. All rights reserved.)

Administrative Data Research Centre England

The Farr Institute London

wellcome trust Fellow

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE