

Challenges and opportunities in using administrative data linkage for research: the importance of quality assessment for understanding bias

Katie Harron and James Doidge

UCL Great Ormond Street Institute of Child Health

January 2020

k.harron@ucl.ac.uk



Record linkage for health data

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.

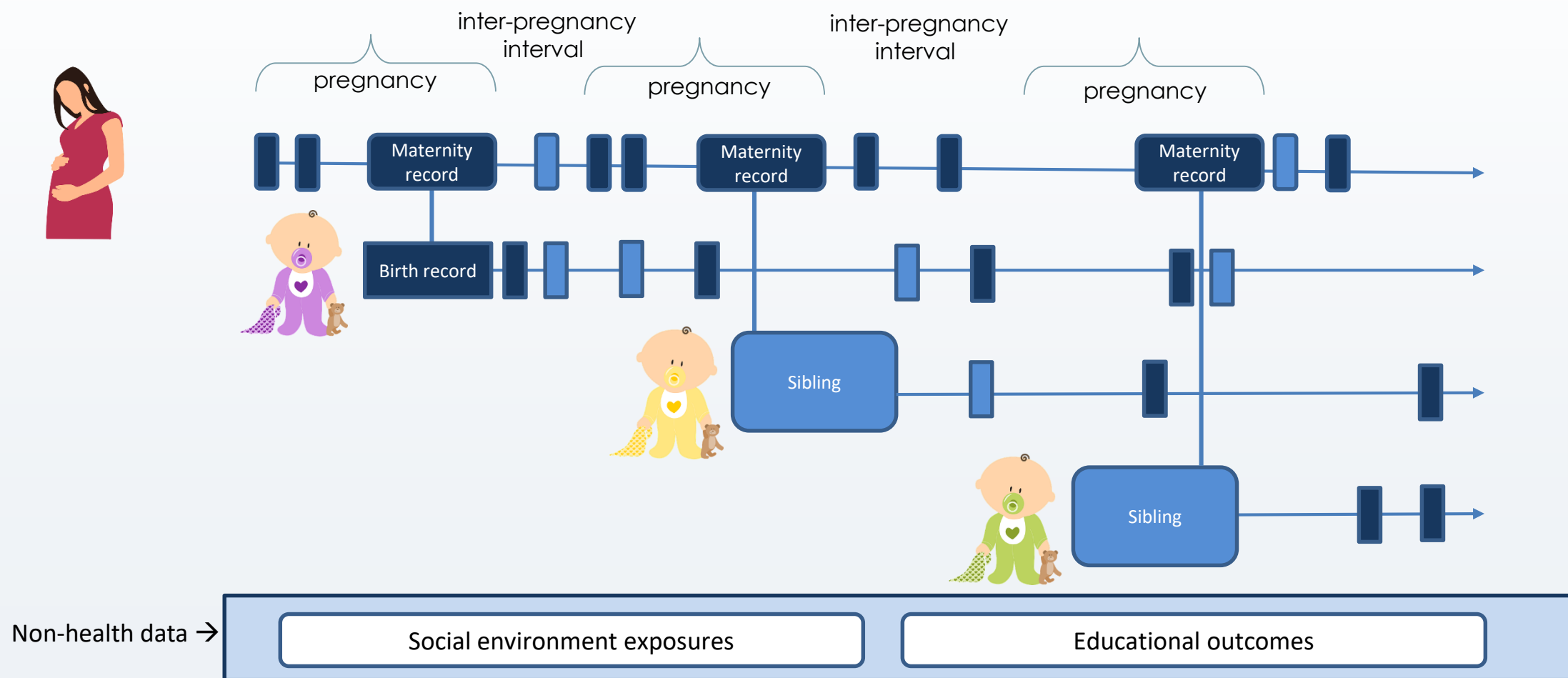
Record linkage is the name given to the process of assembling the pages of this Book, into a volume.

Dunn, 1946

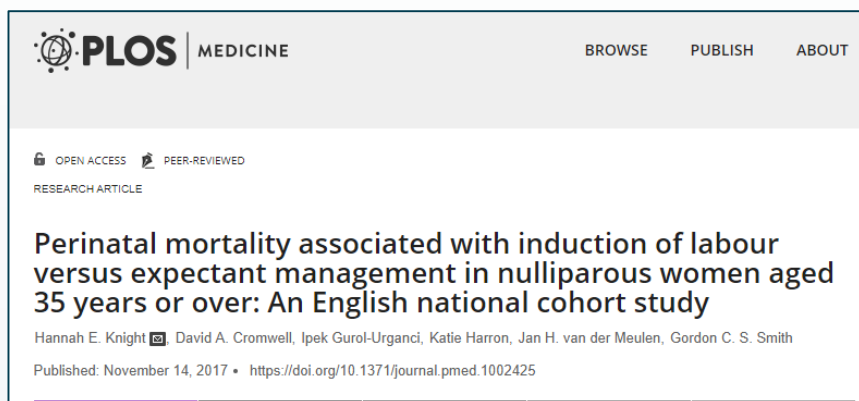


Electronic / administrative data cohorts

- Population cohorts created entirely from linkage of administrative data sources
 - e.g., linkage of mothers and babies within hospital data, in England and beyond...



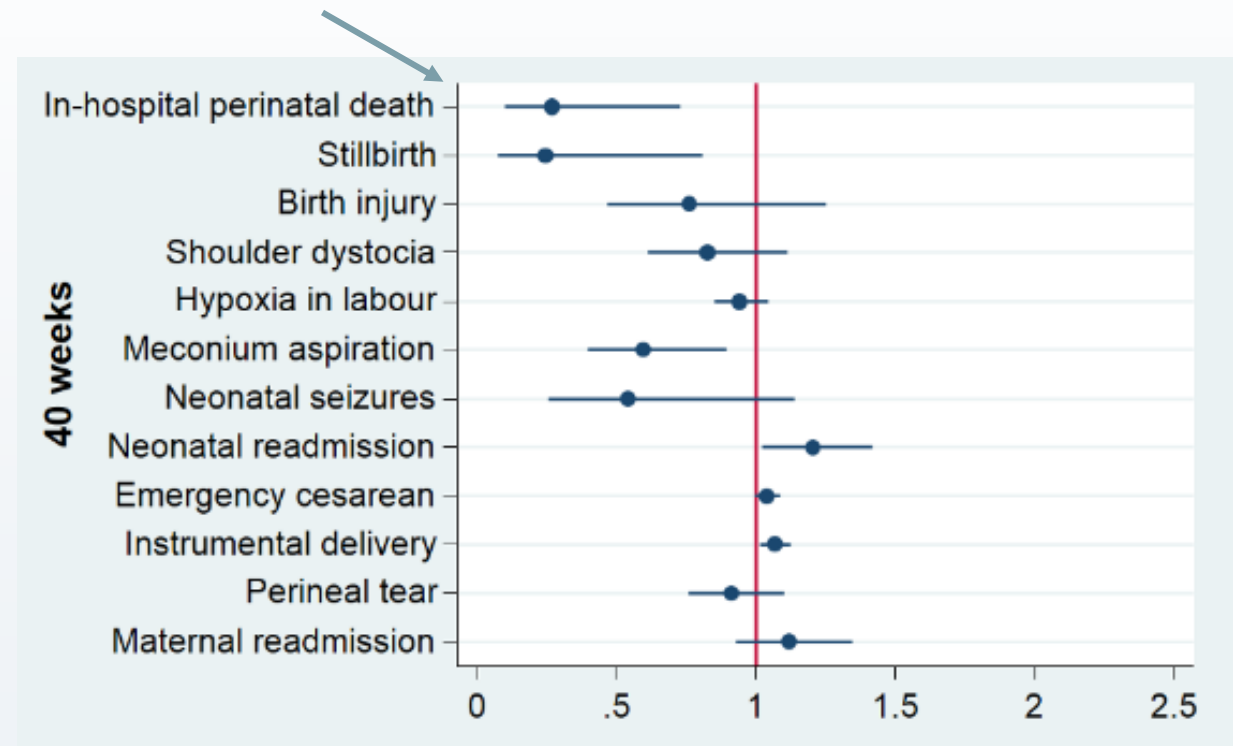
Linking information on perinatal outcomes to maternity records



- RCT evidence suggests induction of labour at 39 weeks has no short-term adverse effect on mother / infant among nulliparous women aged 35 years or older.
- The trial was **underpowered** to address the effect of routine induction of labour on the risk of perinatal death.

66% lower risk of perinatal death (0.08% versus 0.26%)

562 inductions of labour at 40 weeks would be required to prevent 1 perinatal death.



Perinatal outcomes after induction of labour compared with expectant management at 40 weeks gestation

Supplementing mortality data with more complete information on risk factors

Child mortality in England compared with Sweden: a birth cohort study

Ania Zylbersztejn, Ruth Gilbert, Anders Hjerr, Linda Wijlaars, Pia Hardehid

Summary

Background Child mortality is almost twice as high in England compared with Sweden. We aimed to establish the extent to which adverse birth characteristics and socioeconomic factors explain this difference.

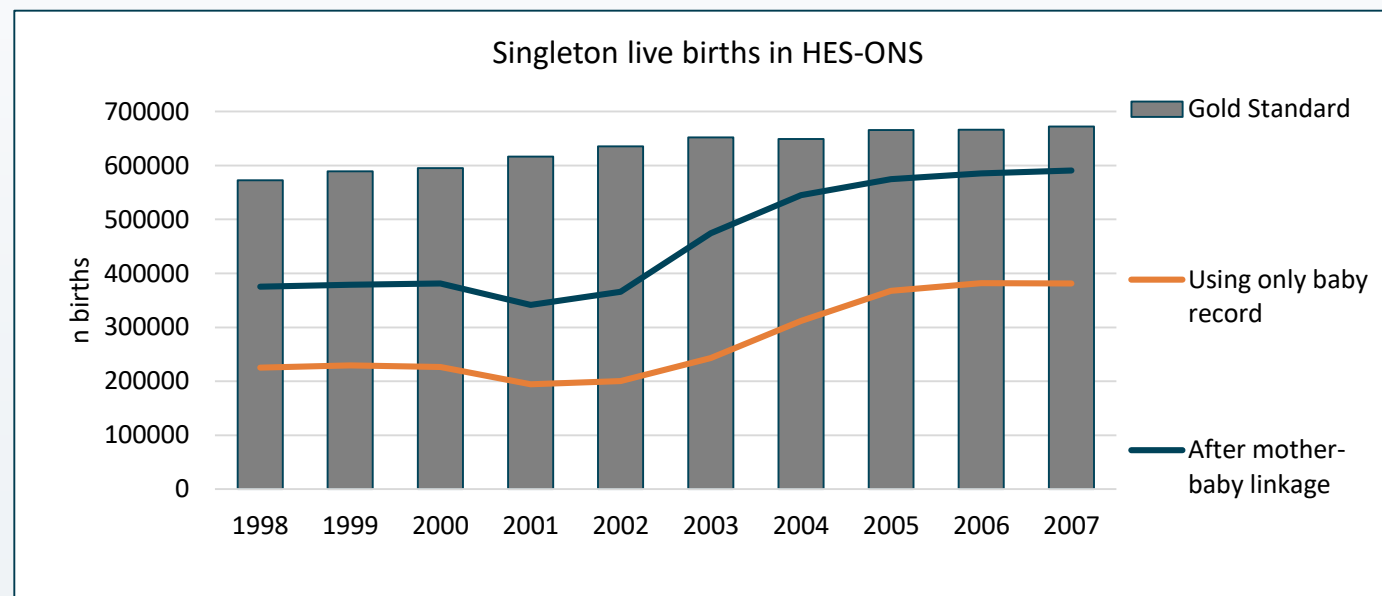
Methods We developed nationally representative cohorts of singleton livebirths between Jan 1, 2003, and Dec 31, 2012

Lancet 391(10134): 2018.

Was excess child mortality in England compared with Sweden explained by the unfavourable distribution of birth characteristics in England?

Linkage to maternal records increased completeness of risk factors:


- 67% to 84% for birth weight
- 64% to 78% for gestational age
- 63% to 97% for maternal age
- 45% to 97% for IMD




The coverage of the complete case cohort increased from **18%** to **75%** of all births in HES-ONS birth cohort.

Challenges

- (Identifier) data quality
- Linkage errors

- 
- Administrative data not designed for linkage
 - **Unique identifiers** may not be present in all sources
 - Requires appropriate **linkage methods**

- 
- False matches and missed matches
 - Can lead to **biased results**
 - Requires **appropriate analysis** methods

How is linkage done?

- Deterministic (rule-based)

1

- Sex
- Date of Birth
- NHS Number

2

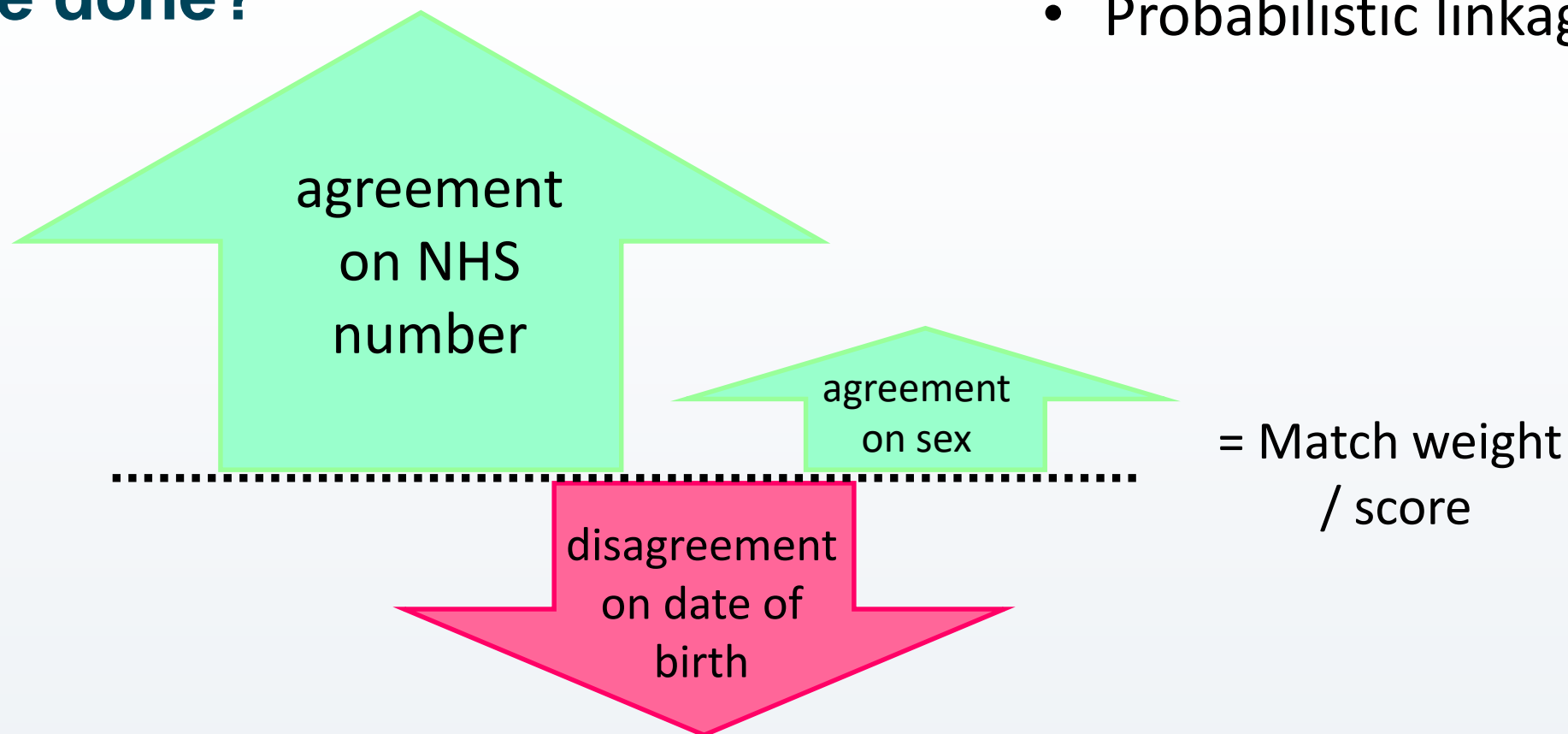
- Sex
- Date of Birth
- Postcode
- Local Patient Identifier within Provider

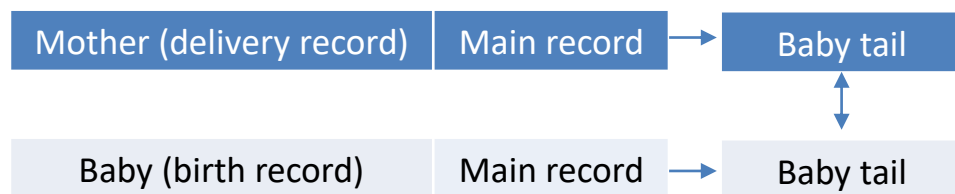
3

- Sex
- Date of Birth
- Postcode

How is linkage done?

- Probabilistic linkage





A	GP practice
B	Postcode district
C	Estimated delivery date
D	First antenatal assessment
E	Episode end
F	Birth weight
G	Episode start
H	Delivery place (Intention)
I	Status of person conducting delivery
J	Maternal age
K	Ethnic group
L	Gestation at first antenatal visit
M	Gestational age
N	Anaesthetic during delivery
O	Method of delivery
P	Method to induce labour
Q	Anaesthetic post-delivery
R	Sex
S	Delivery place
T	Resuscitation method
U	Birth status
V	Number of babies
W	Birth order

 PUBLISH ABOUT

OPEN ACCESS PEER-REVIEWED

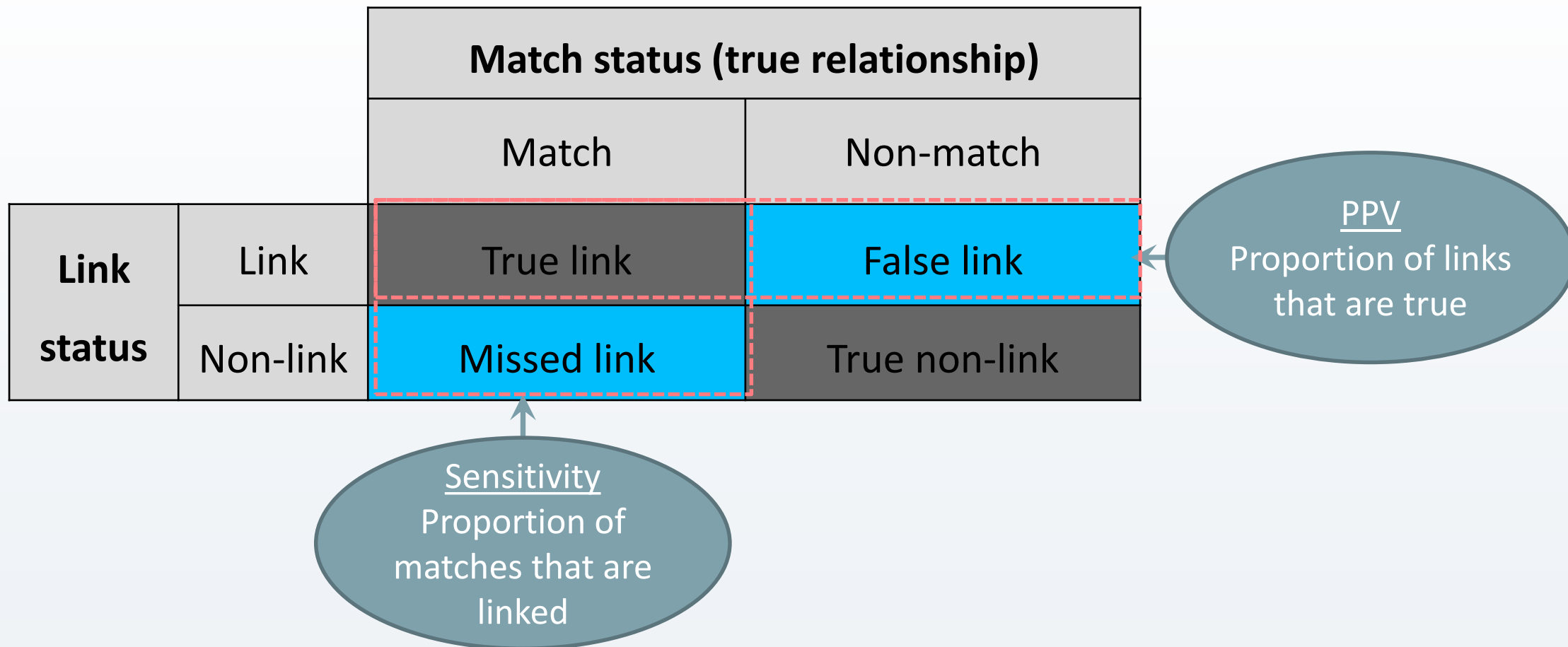
RESEARCH ARTICLE

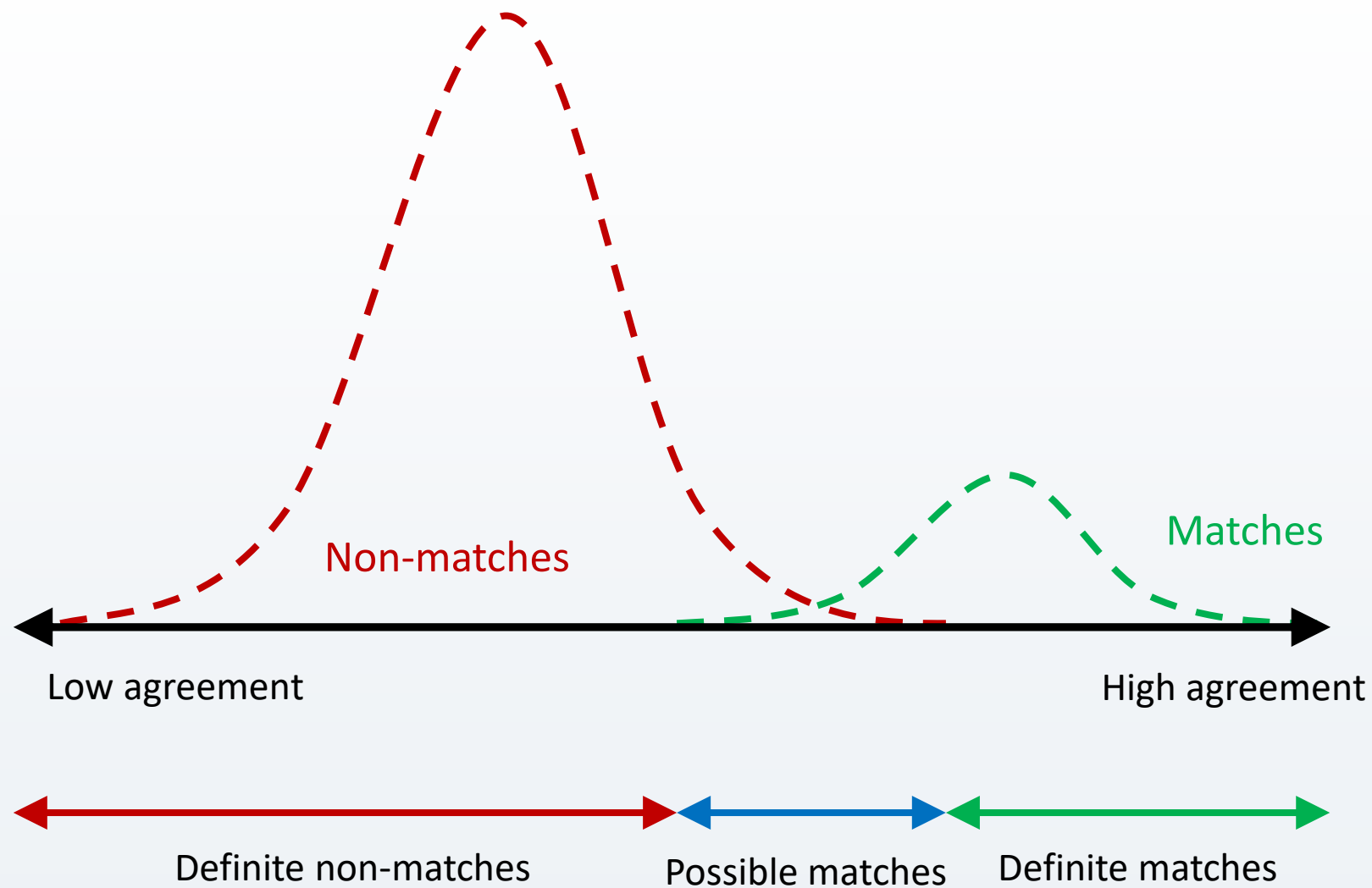
Linking Data for Mothers and Babies in De-Identified Electronic Health Data

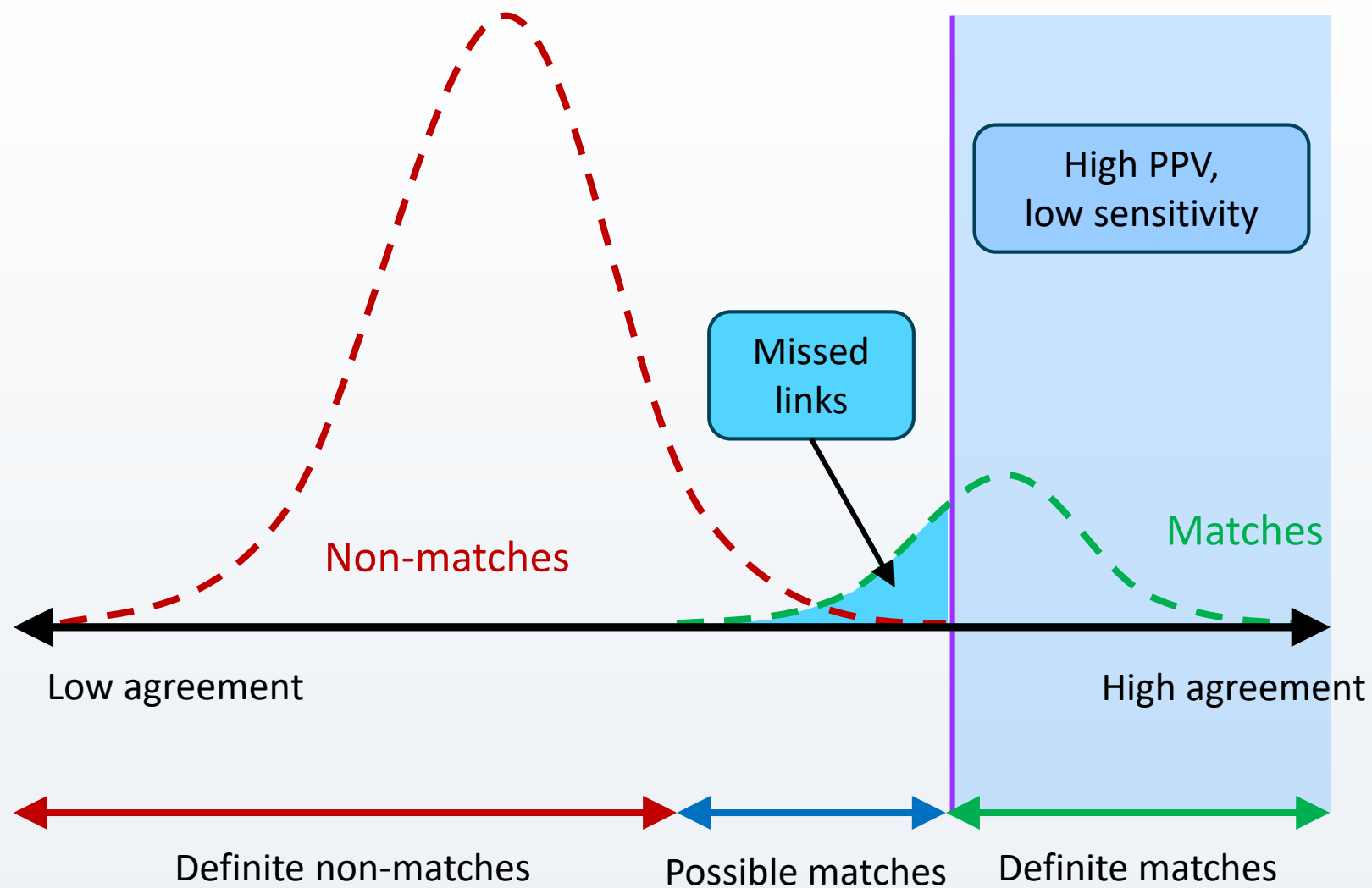
Katie Harron , Ruth Gilbert, David Cromwell, Jan van der Meulen

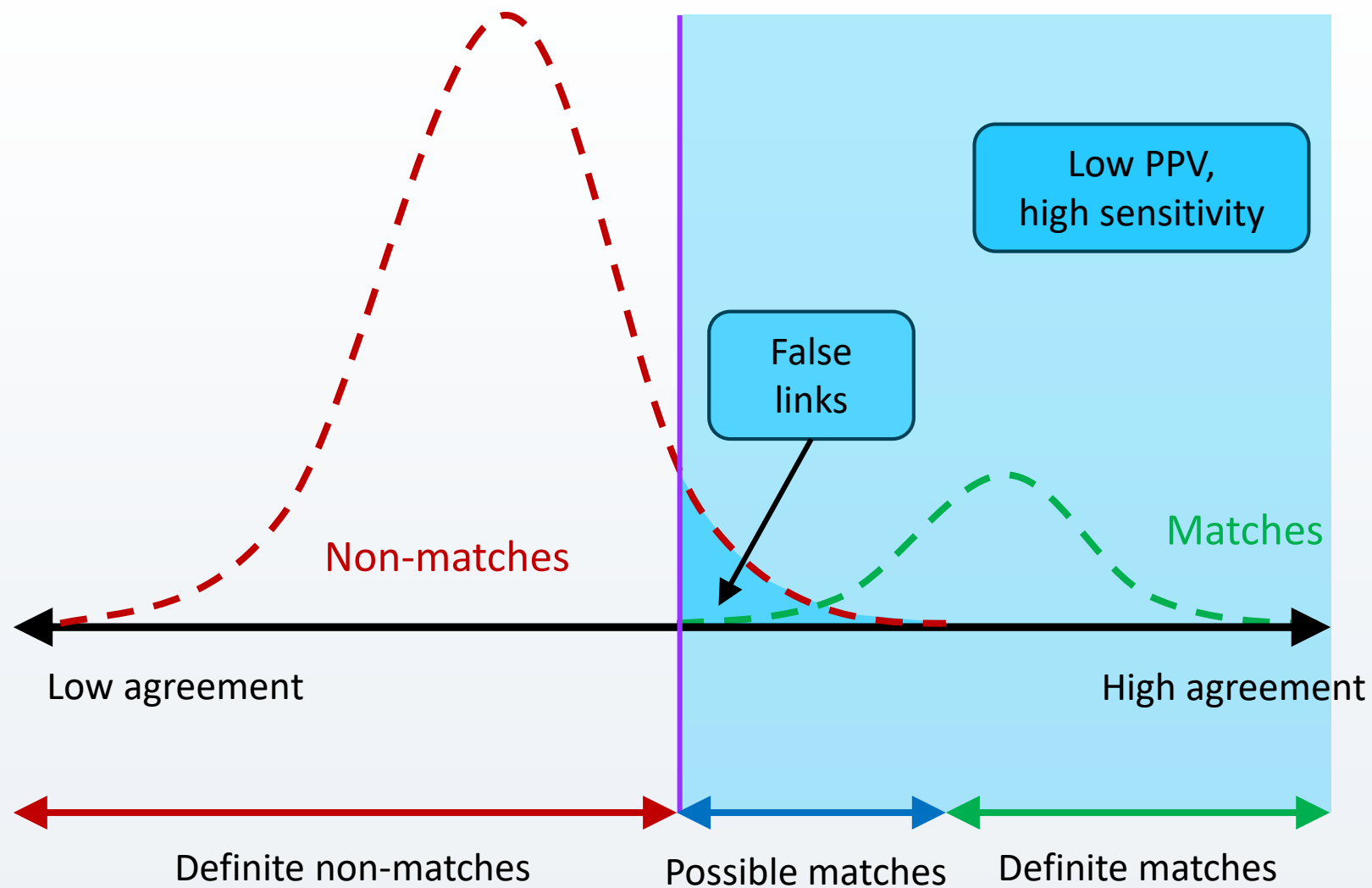
Published: October 20, 2016 • <https://doi.org/10.1371/journal.pone.0164667>

Linkage error









Bias due to linkage error

- Large body of evidence showing that even small amounts of linkage errors can introduce substantial bias to results
 - Particularly important when errors are non-random, or more likely to occur for particular subgroups
- Methods for handling bias due to linkage error have been highlighted as a priority for research

■ Methods to improve data linkage and analysis of complex data. Research methods are needed to underpin efficient linkage of multiple datasets, to quantify potential biases resulting from linkage and how this impacts on results, and to handle linkage error in data analyses. A study to investigate linkage procedures and management in different countries would be very informative.



public health research & practice Sep

Perspective

Routinely collected data as a strategic resource for research: priorities for methods and workforce

Louisa Jorm^{a,b}

^a Centre for Big Data Research in Health, University of New South Wales, Sydney, Australia
^b Corresponding author: l.jorm@unsw.edu.au

Article history

Publication date: September 2015
Citation: Jorm L. Routinely collected data as a strategic resource for research: priorities for methods and workforce. Public Health Res Pract. 2015;25(4):e2541540. doi: http://dx.doi.org/10.17061/php2541540

Abstract

In the era of 'big data', research using routinely collected data offers greater potential than ever before to drive health system effectiveness and efficiency, and population health improvement. In Australia, the policy environment, and emerging frameworks and processes for data governance and access, increasingly support the use of routinely collected data for research. Capitalising on this strategic resource requires investment in both research methods and research workforce.

Priorities for methods development include validation studies, techniques for analysing complex longitudinal data, exploration of bias introduced through linkage error, and a robust toolkit to evaluate policies and programs using 'natural experiments'.

Priorities for workforce development include broadening the skills base of the existing research workforce, and the formation of new, larger, interdisciplinary research teams to incorporate capabilities in computer science, partnership research, research translation and the 'business' aspects of research.

Large-scale, long-term partnership approaches involving government, industry and researchers offer the most promising way to maximise returns on investment in research using routinely collected data.

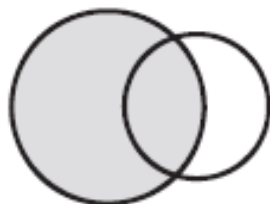
Key points

- Research using routinely collected data can drive health system effectiveness and health improvement
- The policy environment increasingly supports the research use of routinely collected data
- Priorities for methods development include validation studies, and methods for analysing longitudinal data, exploring linkage error, and evaluation using 'natural experiments'

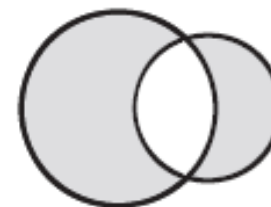
Classifying linkage designs

Impact depends on the linkage classification and the question you are asking...

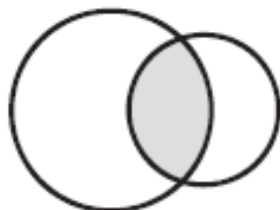
'Master'



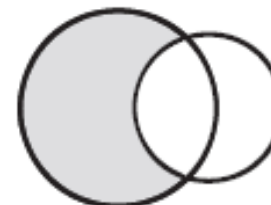
'Disjunctive union'^b



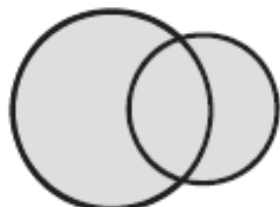
'Intersection'



'Set difference'^b



'Union'



'Perfect overlap'^b



Classifying linkage designs

Impact depends on the linkage classification and the question you are asking...

Question 1: Is the linkage meaningfully interpreted?

Missed matches can lead to...

Yes – to define outcomes (e.g. cancer diagnoses)

- Link / no link = cancer / cancer free

Potential misclassification

Yes – to define study population (e.g. children with Down’s syndrome)

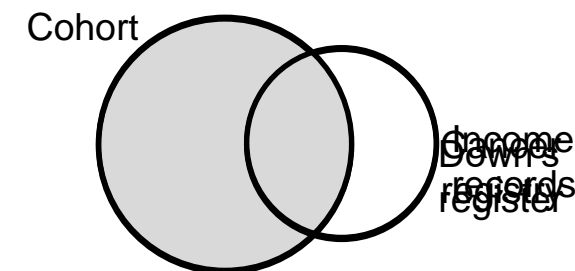
- Link / no link = included / excluded from sample

Potential selection bias / loss of power

No – to add other information (e.g. socio-demographics)

- Link / no link = complete / missing data

False matches can lead to measurement error or misclassification



Impact of missed links



Missing data



Misclassification or measurement error



Erroneous inclusion/exclusion in an analysis



'Splitting' of one person's records into many

	Matched pairs	ISC residuals	MDC residuals
Maternal factors	<i>n</i> = 250 186	<i>n</i> = 2596	<i>n</i> = 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	<i>n</i> = 253 538	<i>n</i> = 1570	<i>n</i> = 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

Impact of missed links



Missing data



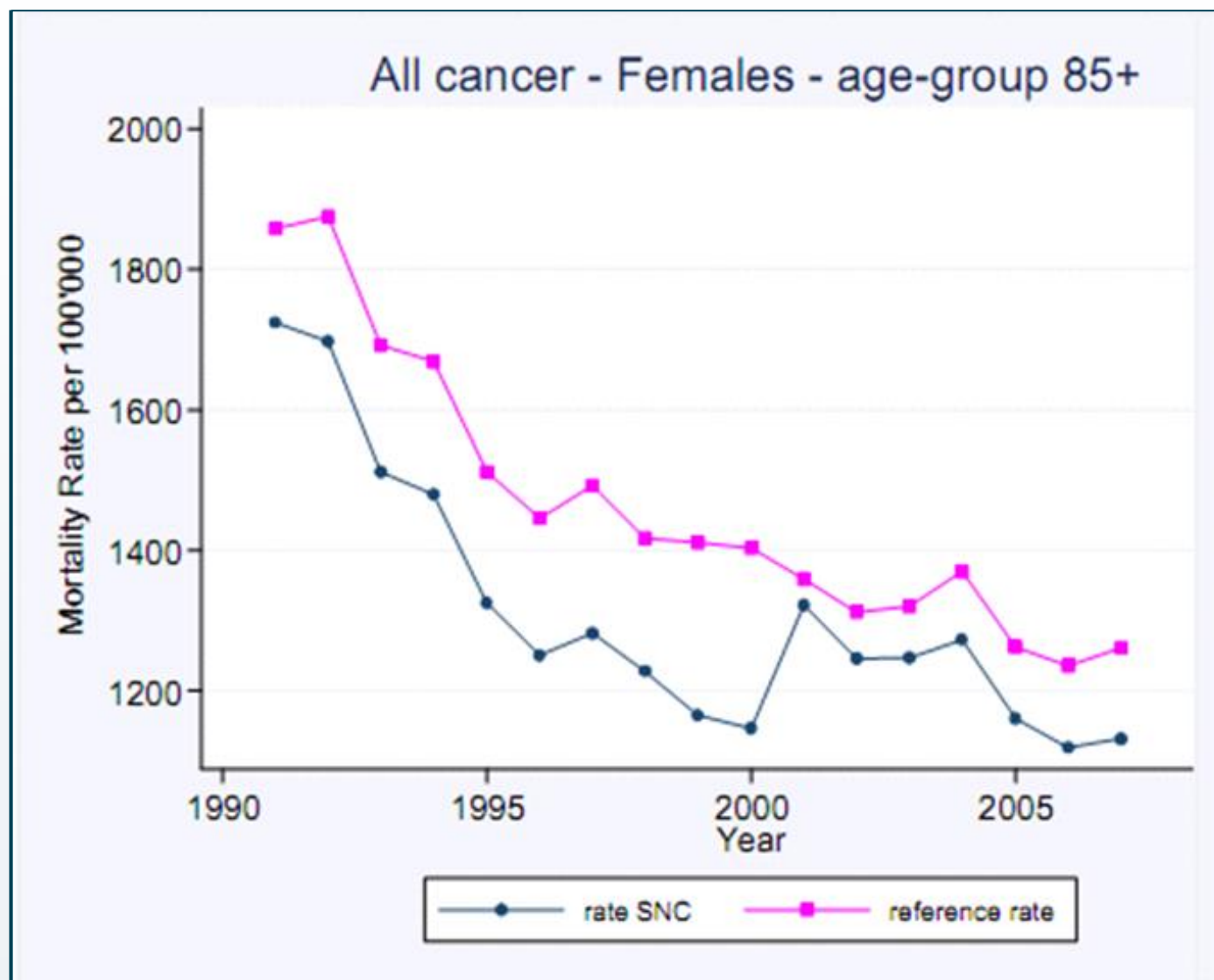
Misclassification and measurement error



Erroneous inclusion/exclusion in an analysis



'Splitting' of one person's records into many



Impact of missed links



Missing data



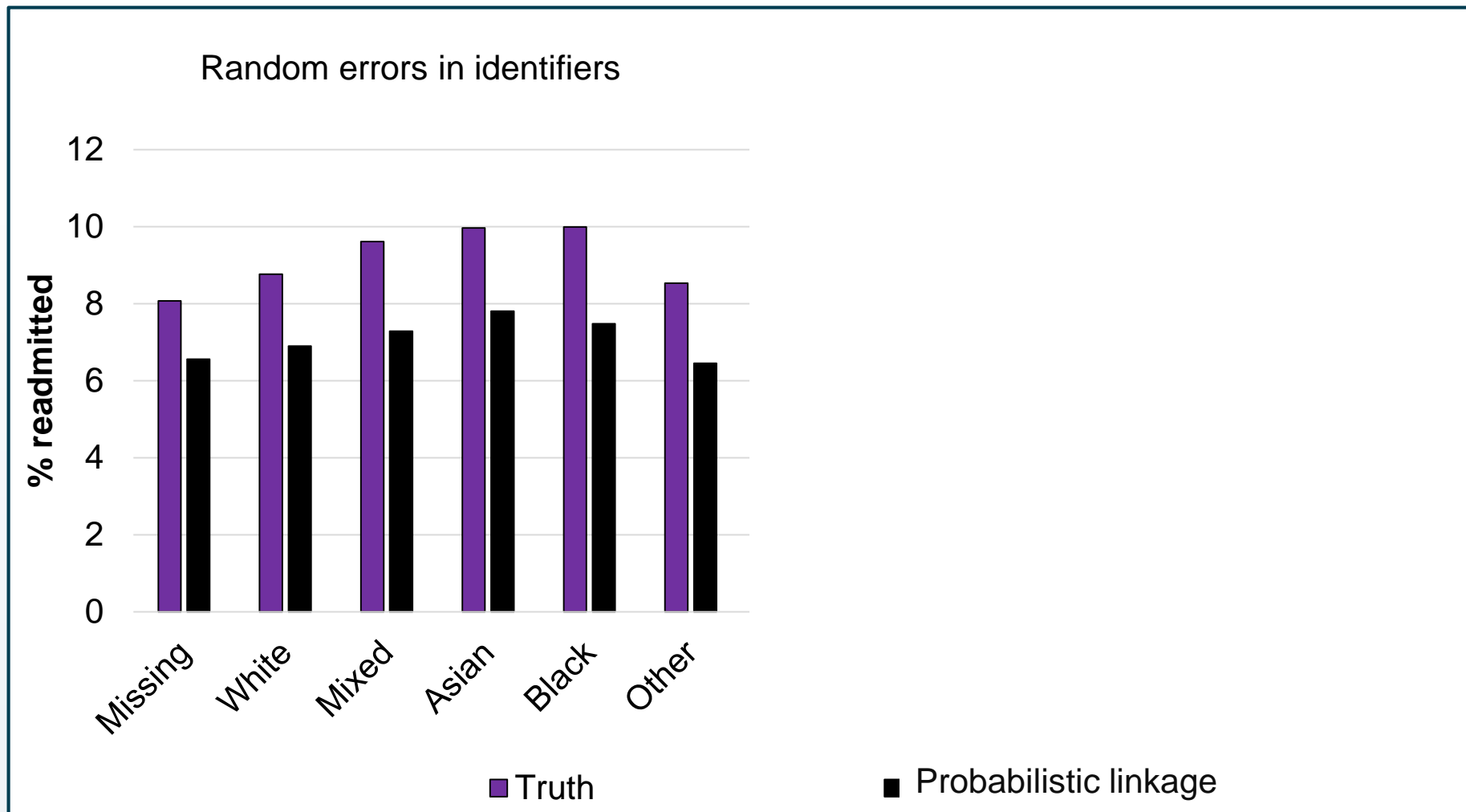
Misclassification and measurement error



Erroneous inclusion/exclusion in an analysis



'Splitting' of one person's records into many



Impact of false links



Misclassification or measurement error



Erroneous inclusion/exclusion in an analysis



'Merging' of multiple people's records into one

Highly sensitive
Highly specific

	Relaxed	NCHS cut-points	Tightened
Table 3. Hazard Ratios for the Association Between Ethnicity and Mortality Using Three Linkage Criteria, 1989-2002			
Ethnicity and nativity			
FB Hispanic	1.24***	0.97	0.78***
US NH White	ref	ref	ref
		* $p < .10$. ** $p < .05$. *** $p < .001$	

Solutions: evaluating linkage error



Gold standard data

- Positive / negative controls
- Comparisons with external references



Comparisons of linked / unlinked records

- Or of high / low quality records



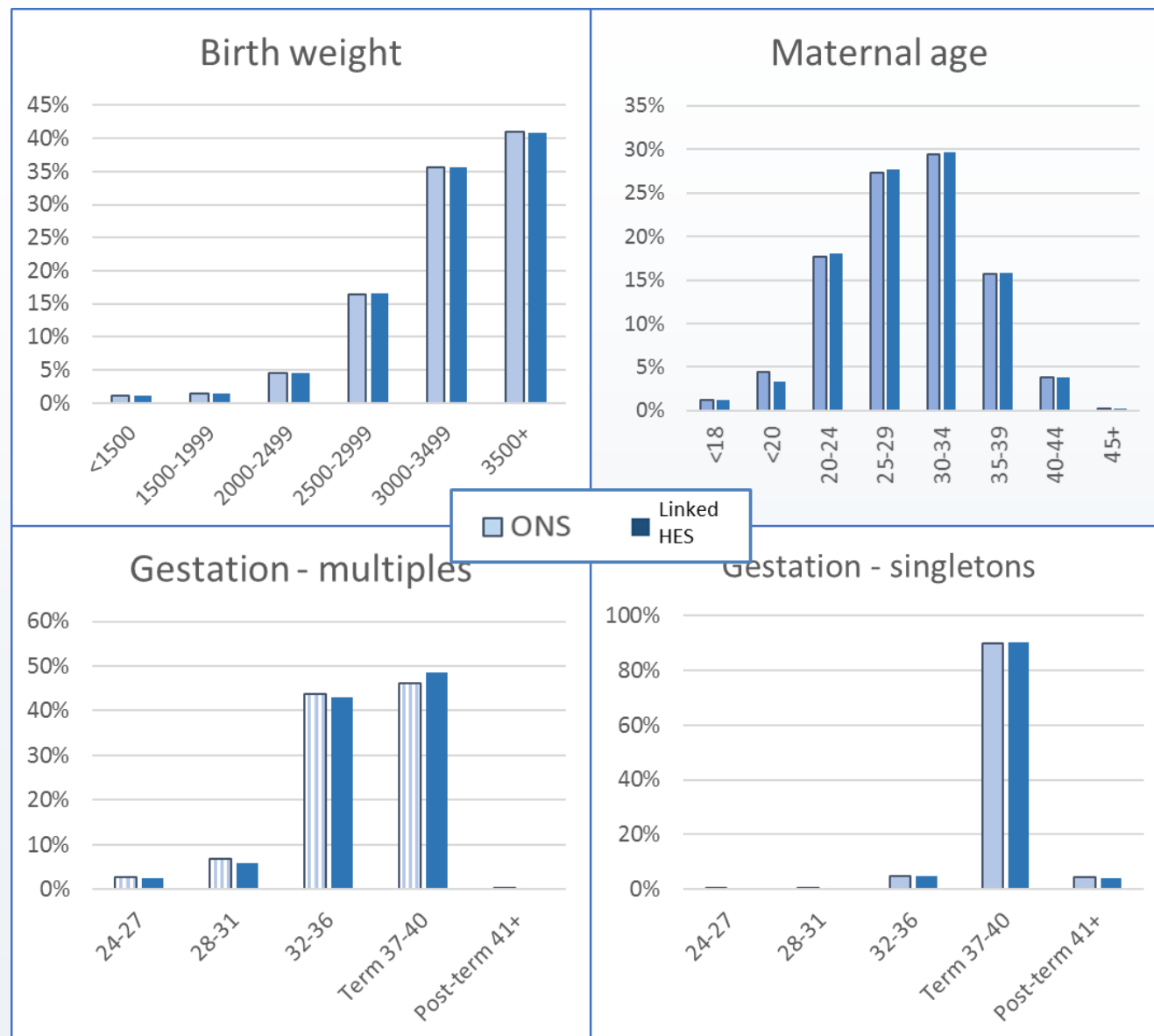
Quality control checks

- Implausible scenarios

Positive / negative controls

- Linking mortality records for prisoners known to still be alive (-)
 - Moore CL et al. 2014. A new method for assessing how sensitivity and specificity of linkage studies affects estimation. *PLoS One*. 9(7):e103690.
- Linking birth registrations for pregnancies known to have an abortive outcome (-)
 - Paixão ES et al. 2019. Validating linkage of multiple population-based administrative databases in Brazil. *PloS One*. 14(3):e0214050-e0214050
- Linking infection surveillance records for neonates with a clinical recording of infection in their admission record (+)
 - Fraser C et al. Linking surveillance and clinical data for evaluating trends in bloodstream infection rates in neonatal units in England. *Submitted*.

Comparisons with external reference data



Harron K, Gilbert R, Cromwell D and van der Meulen J. Linking Data for Mothers and Babies in De-Identified Electronic Health Data. *PLoS One*. 2016; 11: e0164667.
<http://doi.org/10.1371/journal.pone.0164667>

Solutions: evaluating linkage error



Gold standard data

- Positive / negative controls
- Comparisons with external references



Comparisons of linked / unlinked records

- Or of high / low quality records



Quality control checks

- Implausible scenarios

High / low quality records

	All	NHS Number				p-value [†]
		Available and valid		Not available or invalid		
		N	%	N	%	
All	7538	1759	23.3	5779	76.7	
Age group in years						
0 to 14	122	40	32.8	82	67.2	
15 to 44	4724	990	21.0	3734	79.0	
45 to 64	1576	409	26.0	1167	74.0	
65 and over	1061	320	30.2	741	69.8	<0.001
Missing**	55	0	0	55	100.0	
Sex of case						
Female	2941	726	24.7	2215	75.3	
Male	4355	1012	23.2	3343	76.8	
Missing	242	21	8.7	221	91.3	0.15

Solutions: evaluating linkage error



Gold standard data

- Positive / negative controls
- Comparisons with external references



Comparisons of linked / unlinked records

- Or of high / low quality records



Quality control checks

- Implausible scenarios

Quality control checks

- Use evidence that two records do not belong to the same person to identify false-matches
- Admission following death
- Linkage of prostate cancer records with female hospital records

	Infants (<i>n</i> = 733,770)		<i>p</i>
	Not (<i>n</i> = 773,446)	Simultaneous Admission (<i>N</i> = 324)	
Male	51.7%	56.8%	.07
Preterm ^a	7.9%	15.1%	<.001
White ^a	75.8%	66.8%	(ref)
Mixed ^a	4.6%	6.0%	.09
Asian ^a	11.1%	18.4%	<.001
Black ^a	5.3%	4.4%	.83
Chinese ^a	0.6%	1.0%	.26
Other ^a	2.7%	3.5%	.22
Multiple birth ^a	3.5%	3.8%	.75

Summary

- Linkage with administrative data is extremely valuable and can be more efficient than traditional follow-up
 - Cohorts created entirely from linked administrative data can provide new resources on a much larger scale than previously possible
- Data quality and linkage errors can challenge the reliability of linked data for analysis
 - Probabilistic linkage methods can provide measures of certainty
 - Mechanisms for linkage errors can be complex
- Methods for handling linkage errors can lead to more robust research
 - Quantitative bias analysis
 - Imputation-based approaches: Goldstein 2012. *The analysis of record-linked data using multiple imputation with data value priors.* Stat Med 31(28): 3481-3493.

Acknowledgements

Harvey Goldstein, Ruth Gilbert, Jan van der Meulen, James Doidge, Angie Wade, Gareth Hagger-Johnson

Funding:

Wellcome Trust grant numbers 103975/Z/14/Z and 212953/Z/18/Z.

