Using administrative data linkage to create electronic birth cohorts: opportunities and challenges

Katie Harron

UCL Great Ormond Street Institute of Child Health

December 2021

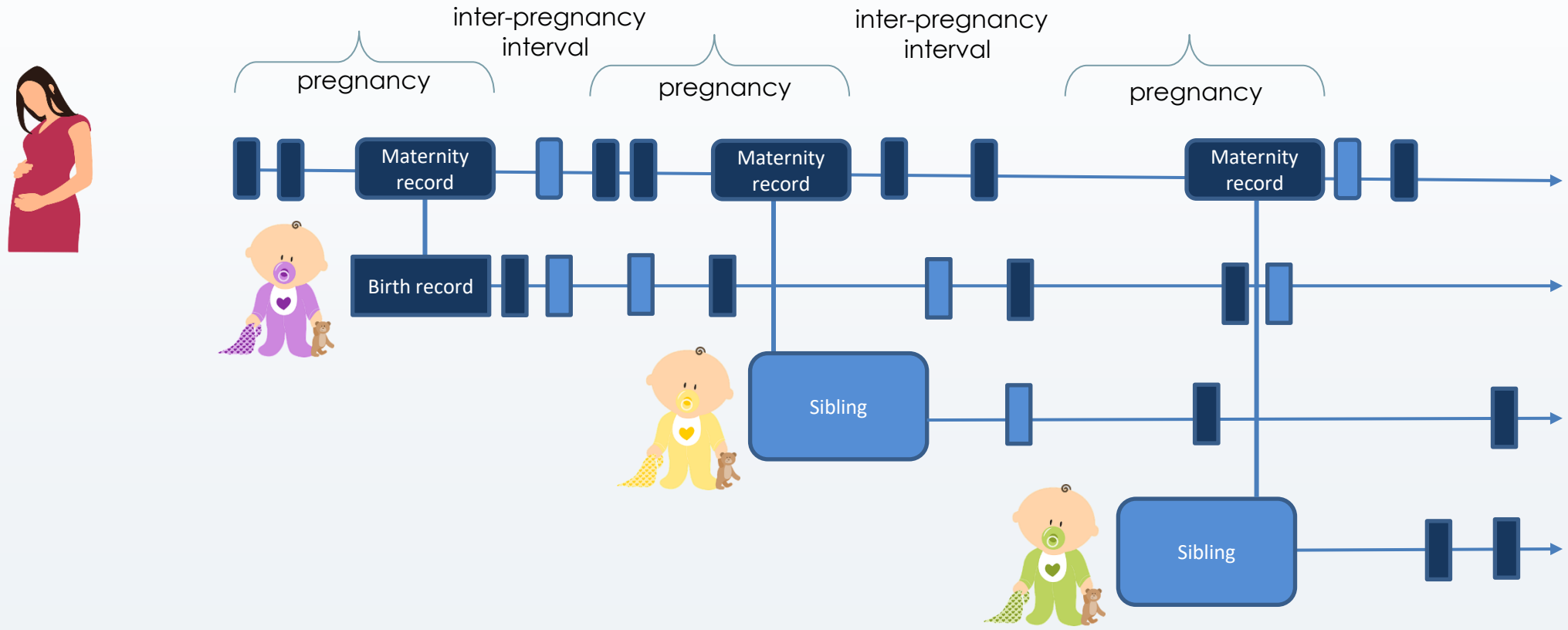
k.harron@ucl.ac.uk

CHIG child health informatics group

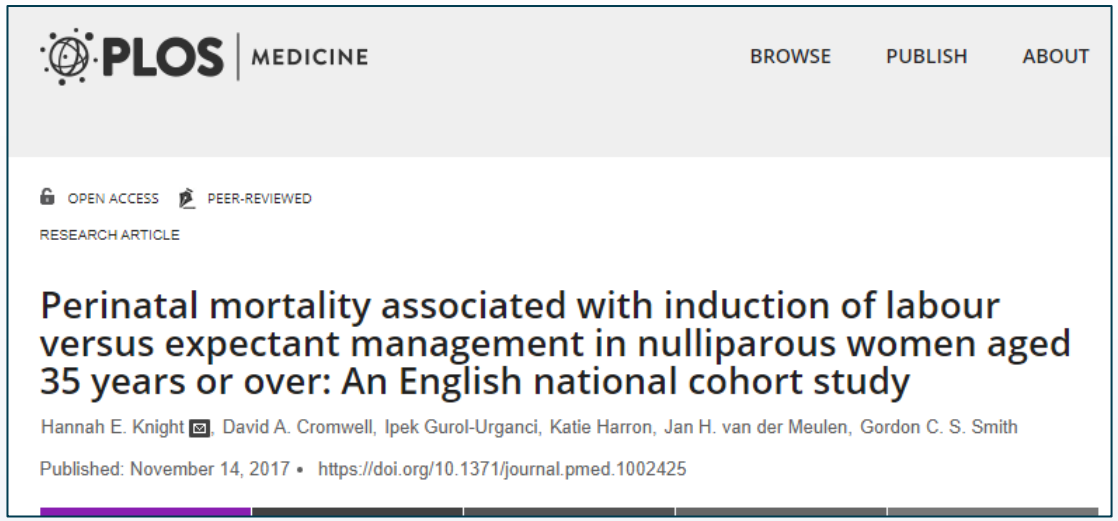
HDRUK Health Data Research UK

Electronic / administrative data cohorts

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



1. Answering important questions that cannot be addressed using traditional approaches

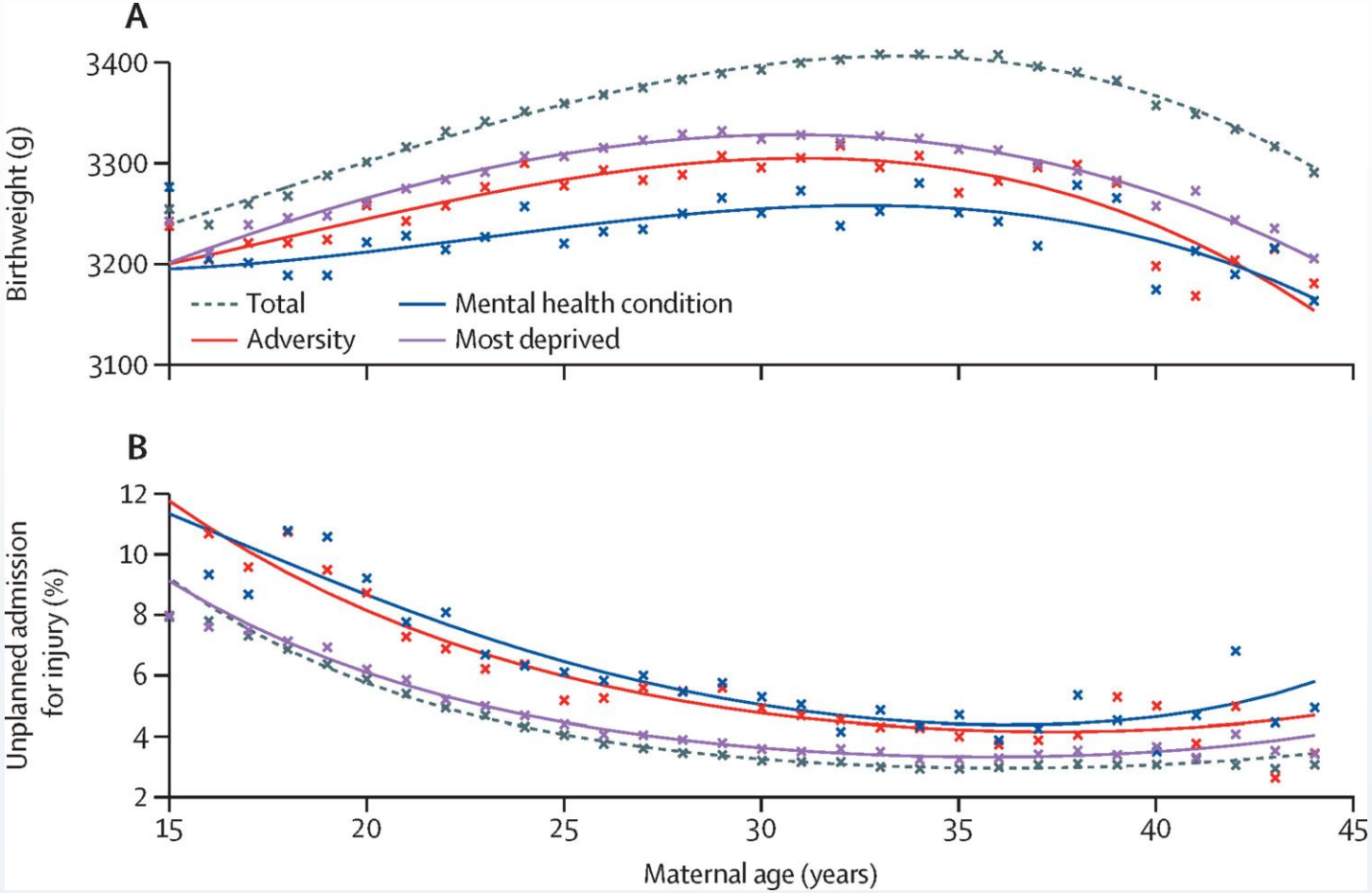


- RCT evidence suggests early induction of labour has no short-term adverse effect on mother / infant among nulliparous women aged 35 years or older.
- The trial was **underpowered** to address the effect of routine induction of labour on the risk of perinatal death.

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

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

2. Identifying early indicators of need



Babies born to mothers with a history of mental health or behavioural conditions were 124g lighter (95% CI 114–134 g) than those born to mothers without these conditions.

For teenage mothers compared with older mothers, 3·6% (95% CI 3·3–3·9%) more infants had an unplanned admission for injury, and there were 10·2 (95% CI 7·5–12·9) more deaths per 10 000 infants.

Harron et al 2021. Associations between pre-pregnancy psychosocial risk factors and infant outcomes: a population-based cohort study in England. The Lancet Public Health. 6(2):e97-e105

3. Improving the quality of administrative data

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

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

Ania Zylbersztejn, Ruth Gilbert, Anders Hjern, Linda Wijlaars, Pia Hardelid

Summary

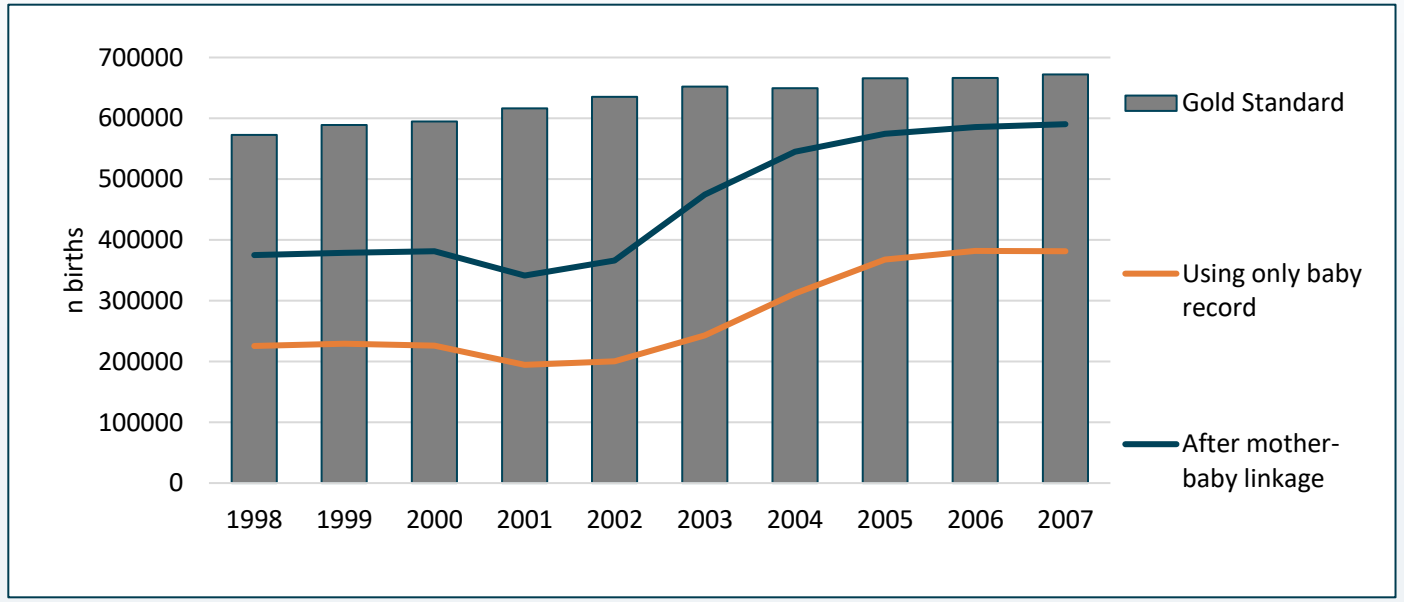
Background Child mortality is almost twice as high in England compared with Sweden. We aimed to establish the extent to which adverse birth characteristics and socioeconomic factors explain this difference.

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

Lancet 391(10134): 2018.

Incomplete recording of risk factors in baby records:

- Birth weight 67% → 84%
- Gestational age 64% → 78%
- Maternal age 63% → 97%
- IMD 45% → 97%

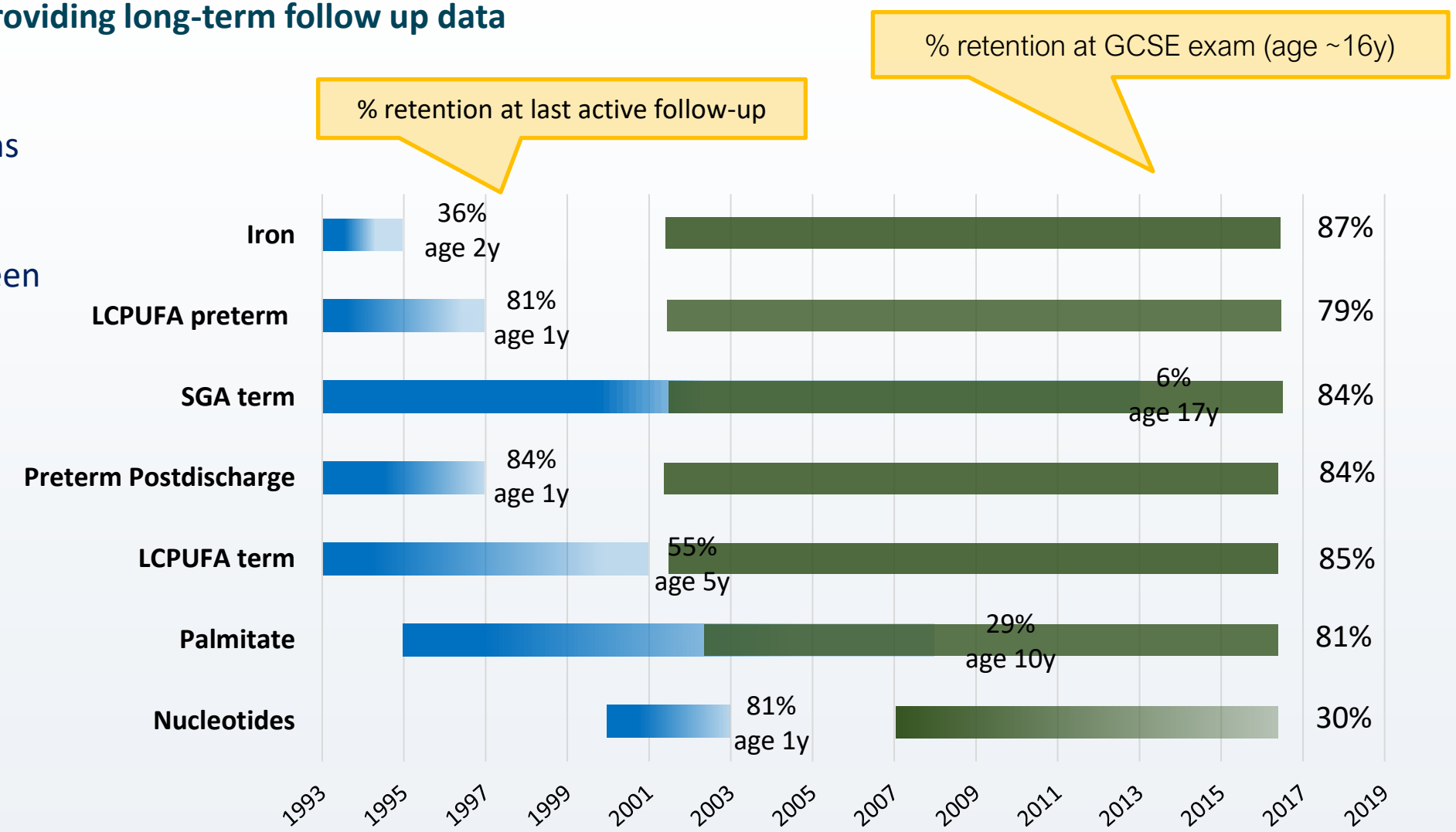


Complete case cohort increased from **18% to 75%** of all births.

Data/figure courtesy of Ania Zylbersztejn

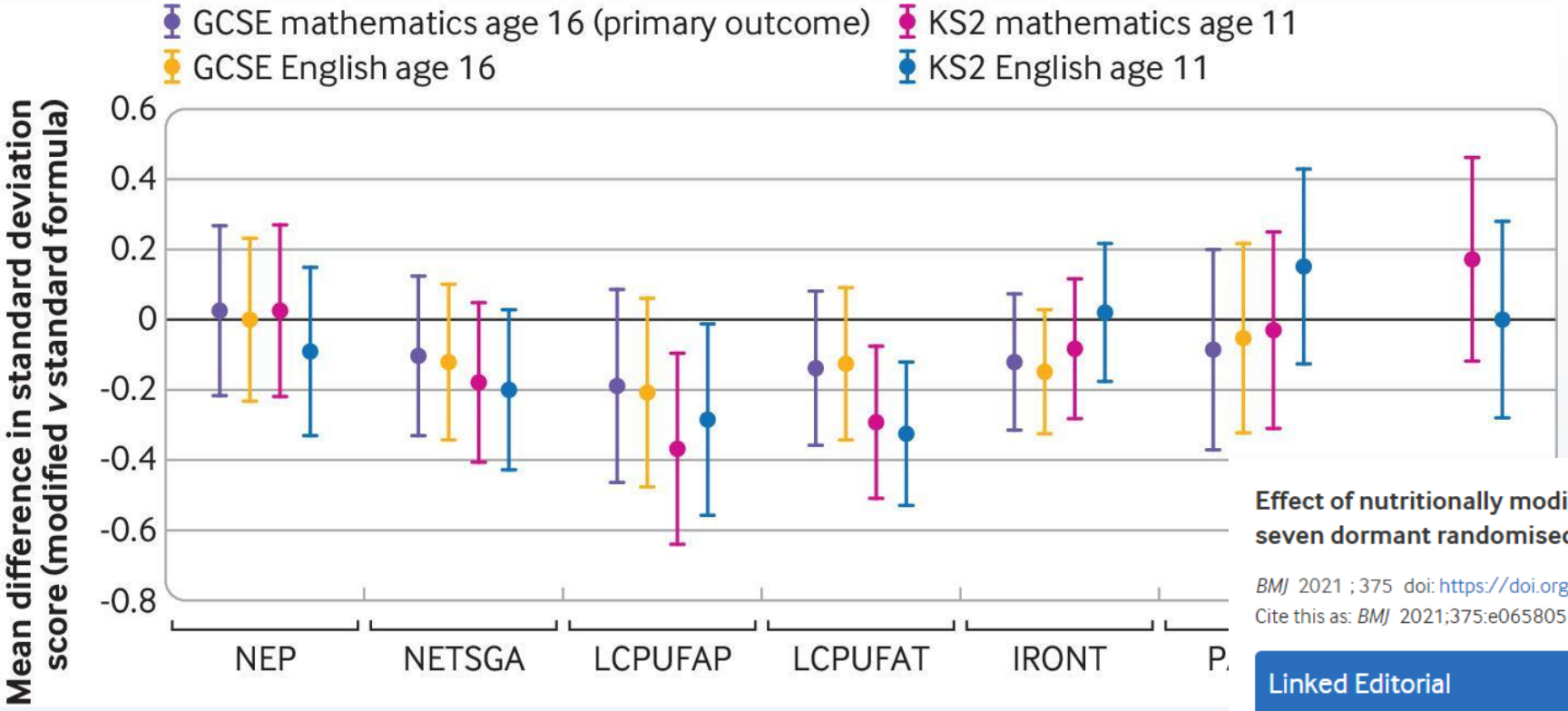
5. Enhancing clinical trials by providing long-term follow up data

- Early nutritional interventions
- 7 infant formula trials
- Conducted in England between 1993-2002
- 2788 participants
- Now aged 17-27 years old



Data/figure courtesy of Maximiliane Verfuerden

5. Enhancing clinical trials by providing long-term follow up data



Effect of nutritionally modified infant formula on academic performance: linkage of seven dormant randomised controlled trials to national education data

BMJ 2021 ; 375 doi: https://doi.org/10.1136/bmj-2021-065805 (Published 11 November 2021)
Cite this as: *BMJ* 2021;375:e065805

Linked Editorial
Enriched formula milks and academic performance in later childhood

Article | Related content | Metrics | Responses | Peer review

Maximiliane L Verfürden, postdoctoral researcher [1], Ruth Gilbert, professor [1], Alan Lucas, professor [1], John Jerrim, professor [2], Mary Fewtrell, professor [1]

Challenges

Privacy / confidentiality →
- Access to linked data is often extremely time consuming
- Researchers typically do not access data in the clear

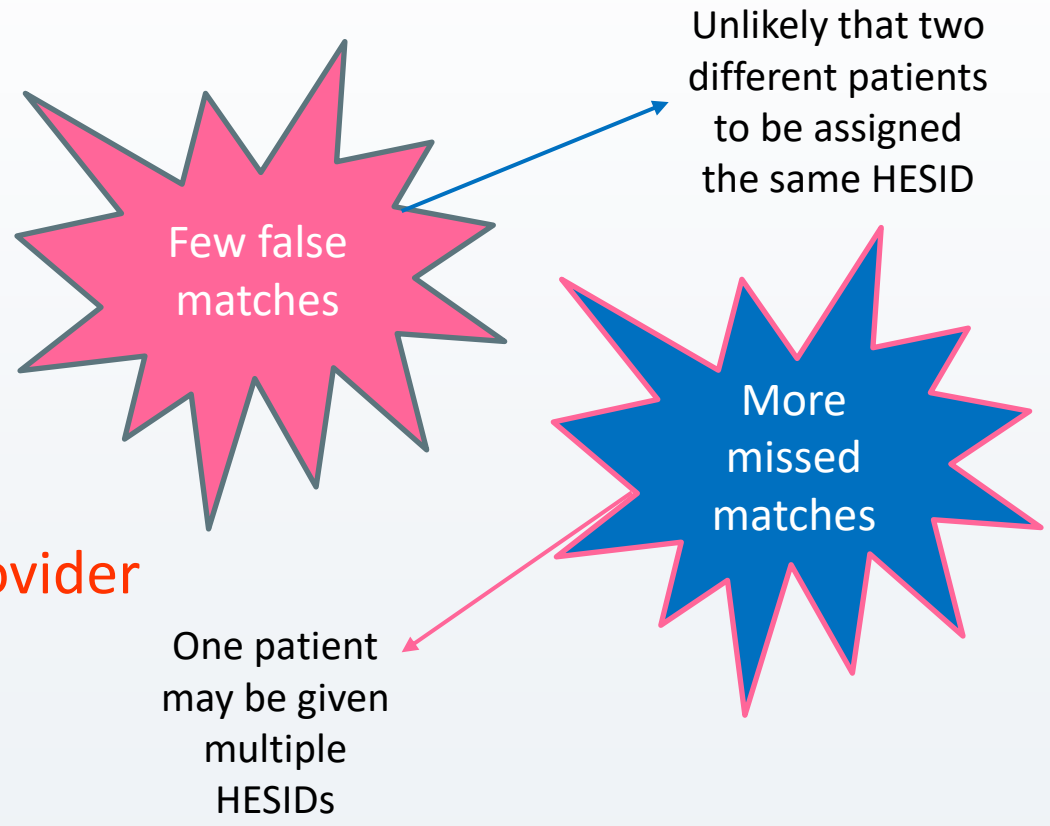
(Identifier) data quality →
- Administrative data not designed for linkage
- Unique identifiers may not be present in all sources
- Requires appropriate linkage methods

Linkage errors →
- False matches and missed matches
- Can lead to biased results
- Requires appropriate analysis methods

How is linkage done?

- Deterministic (rule-based)

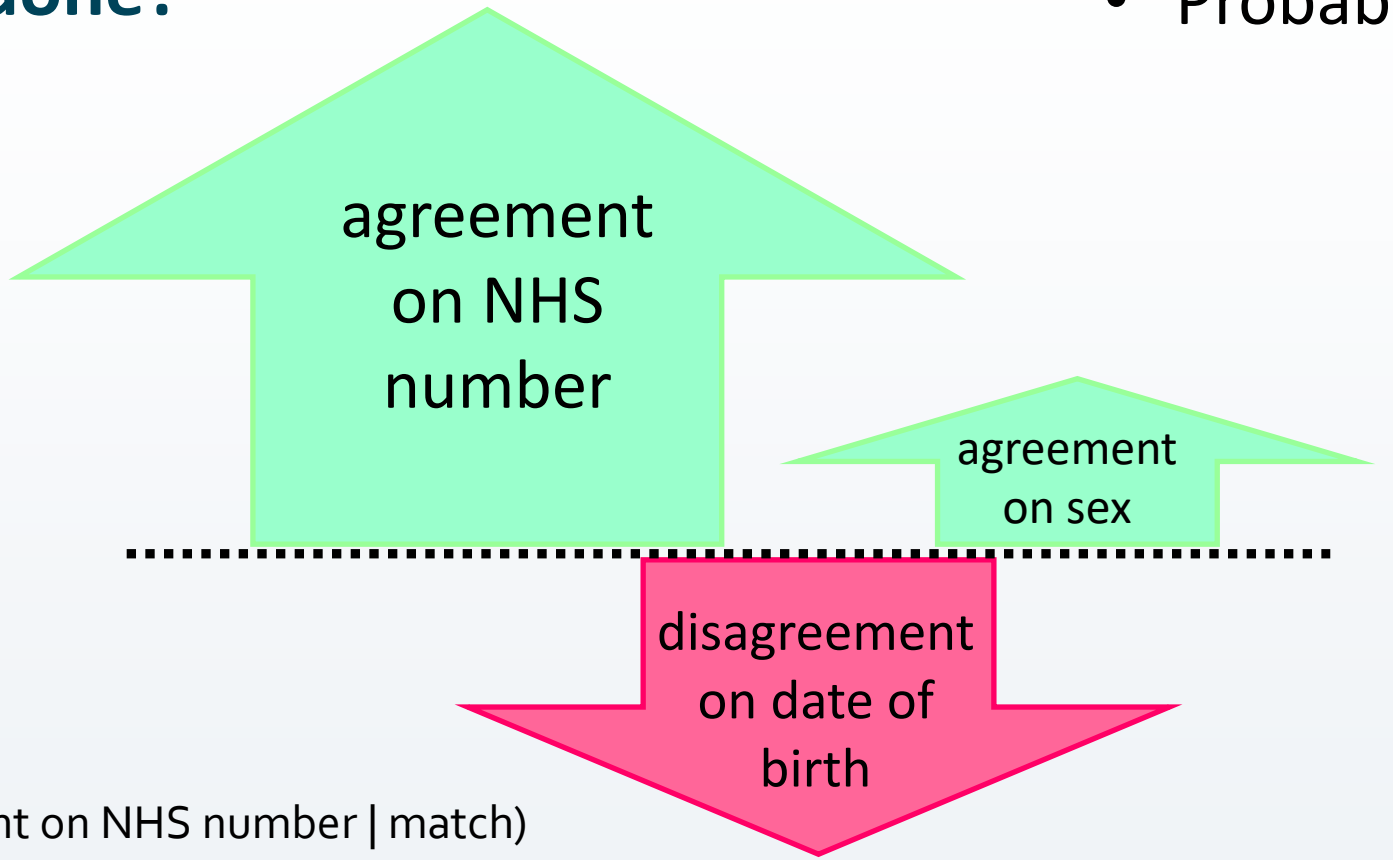
1
   - Sex
   - Date of Birth
   - NHS Number
2
   - Sex
   - Date of Birth
   - Postcode
   - Local Patient Identifier within Provider
3
   - Sex
   - Date of Birth
   - Postcode



Hagger-Johnson, G., et al. (2015). "Data linkage errors in hospital administrative data when applying a pseudonymisation algorithm to paediatric intensive care records." BMJ Open 5(8).

How is linkage done?

- Probabilistic (score based)



= Match weight / score

<u>Fellegi Sunter</u>

m-probability: P(agreement on NHS number | match)

u-probability: P(agreement on NHS number | non-match)

Weight = $\sum \log_2(m/u)$

- Fellegi & Sunter. A theory for record linkage. J Am Stat Assoc. 1969;64(328):1183-210.
- Goldstein et al. A scaling approach to record linkage. Stat Med. 2017;36:2514-21.

Mother (delivery record)	Main record	Baby tail
Baby (birth record)	Main record	Baby tail

PLOS ONE | PUBLISH ABOUT

OPEN ACCESS PEER-REVIEWED

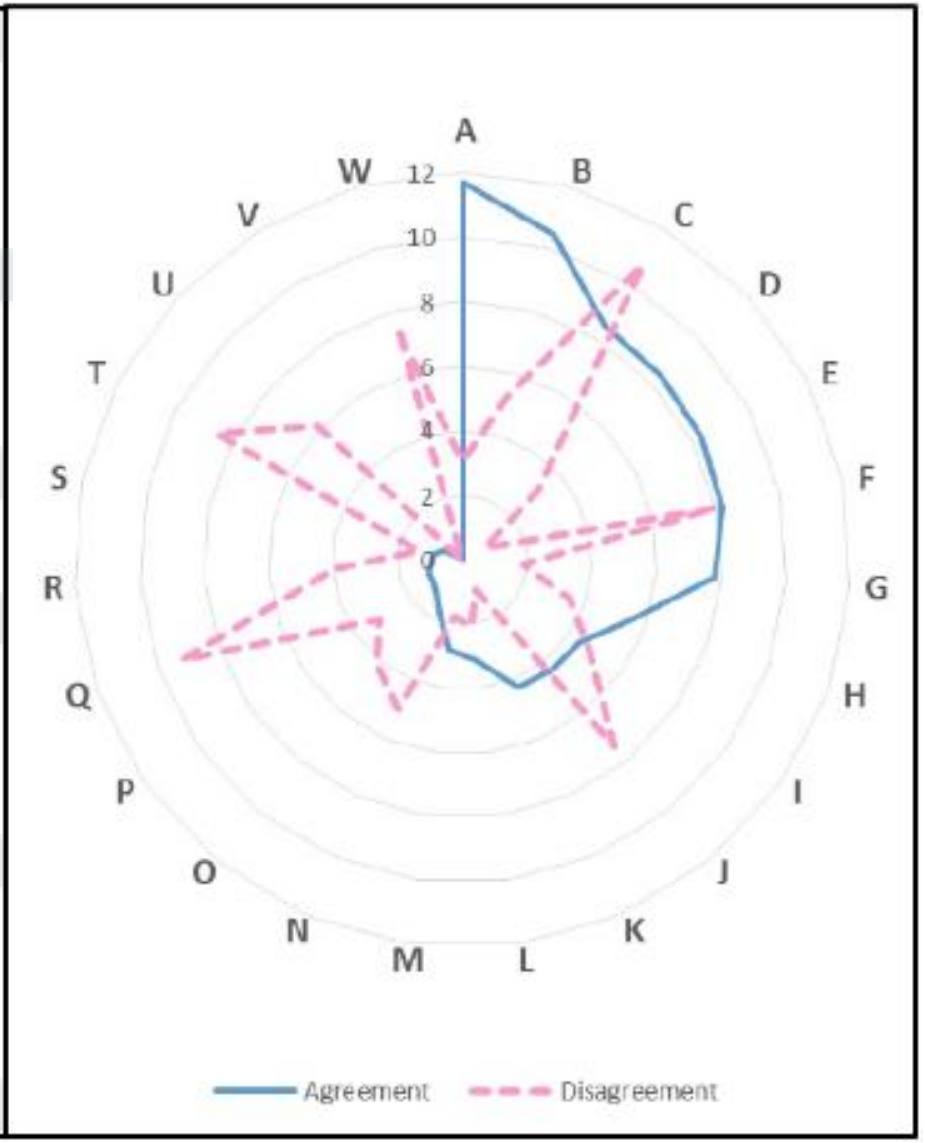
RESEARCH ARTICLE

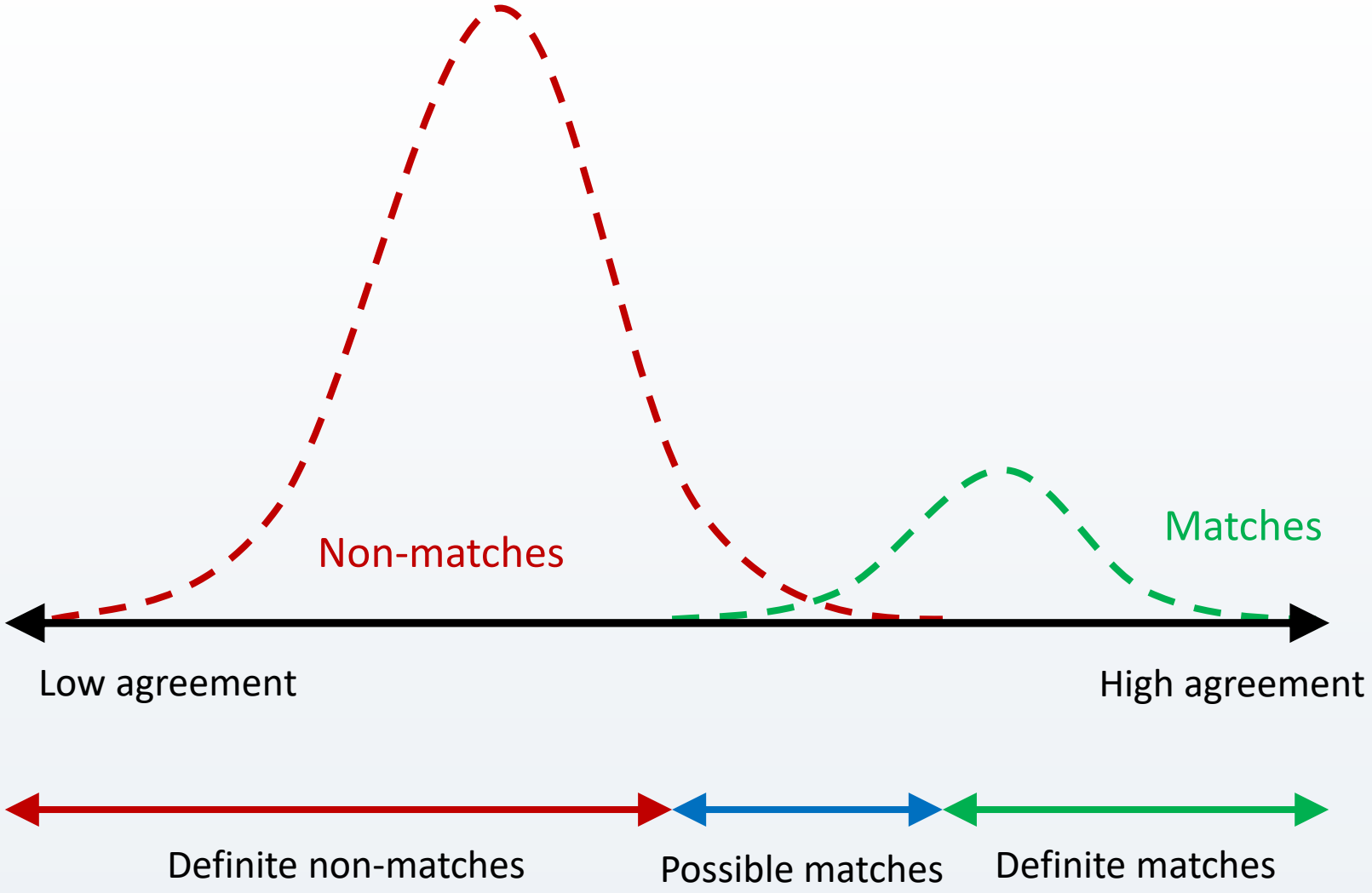
Linking Data for Mothers and Babies in De-Identified Electronic Health Data

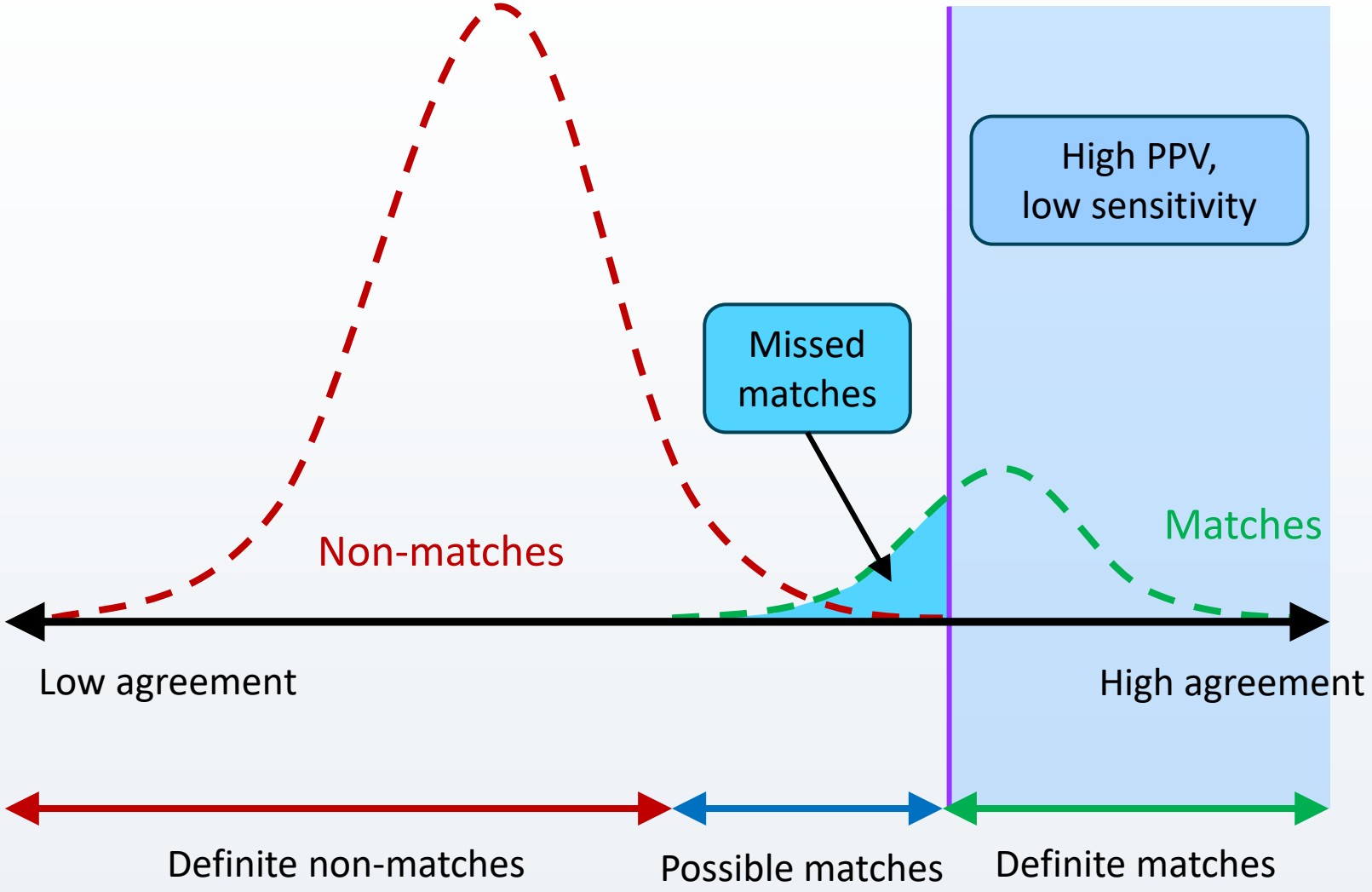
Katie Harron, Ruth Gilbert, David Cromwell, Jan van der Meulen

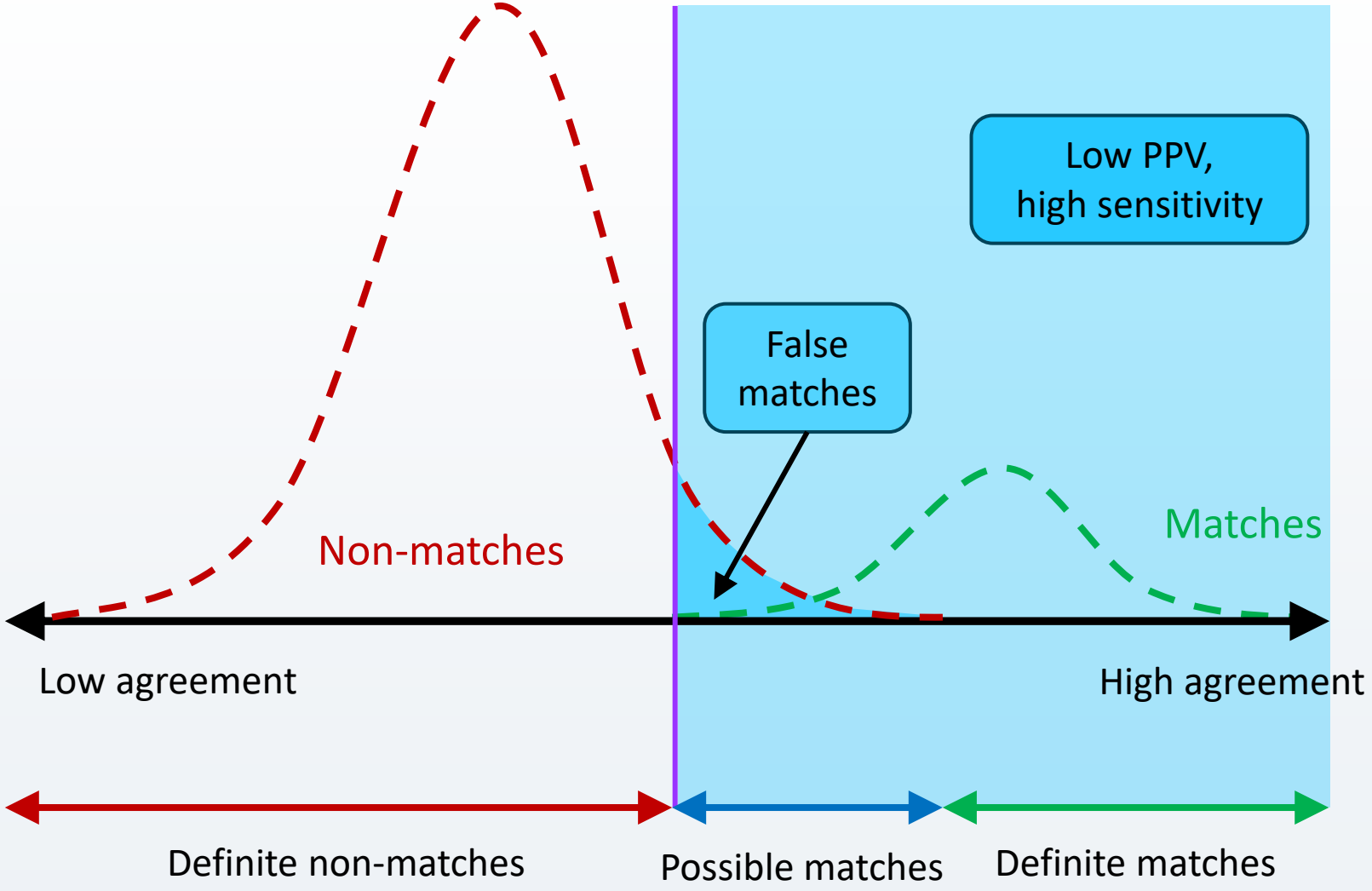
Published: October 20, 2016 • https://doi.org/10.1371/journal.pone.0164667

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




Non-matches
Matches
Low agreement
High agreement
Definite non-matches
Possible matches
Definite matches


High PPV,
low sensitivity
Missed
matches
Non-matches
Matches
Low agreement
High agreement
Definite non-matches
Possible matches
Definite matches

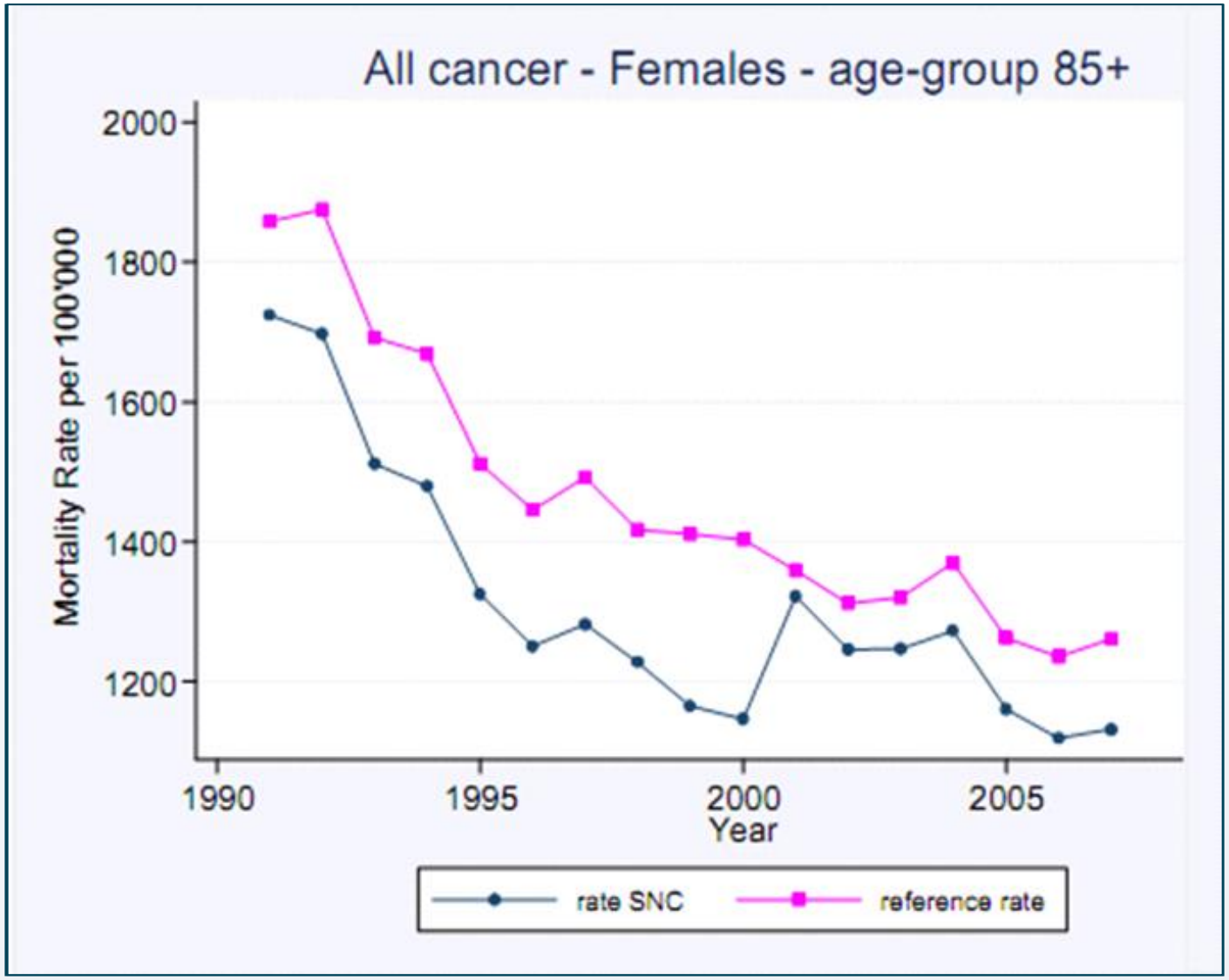

Low PPV,
high sensitivity
False
matches
Non-matches
Matches
Low agreement
High agreement
Definite non-matches
Possible matches
Definite matches

How does linkage error lead to bias?

1: Missed matches

- Missing data
- Misclassification or measurement error
- Erroneous inclusion/exclusion in an analysis
- ‘Splitting’ of one person’s records into many

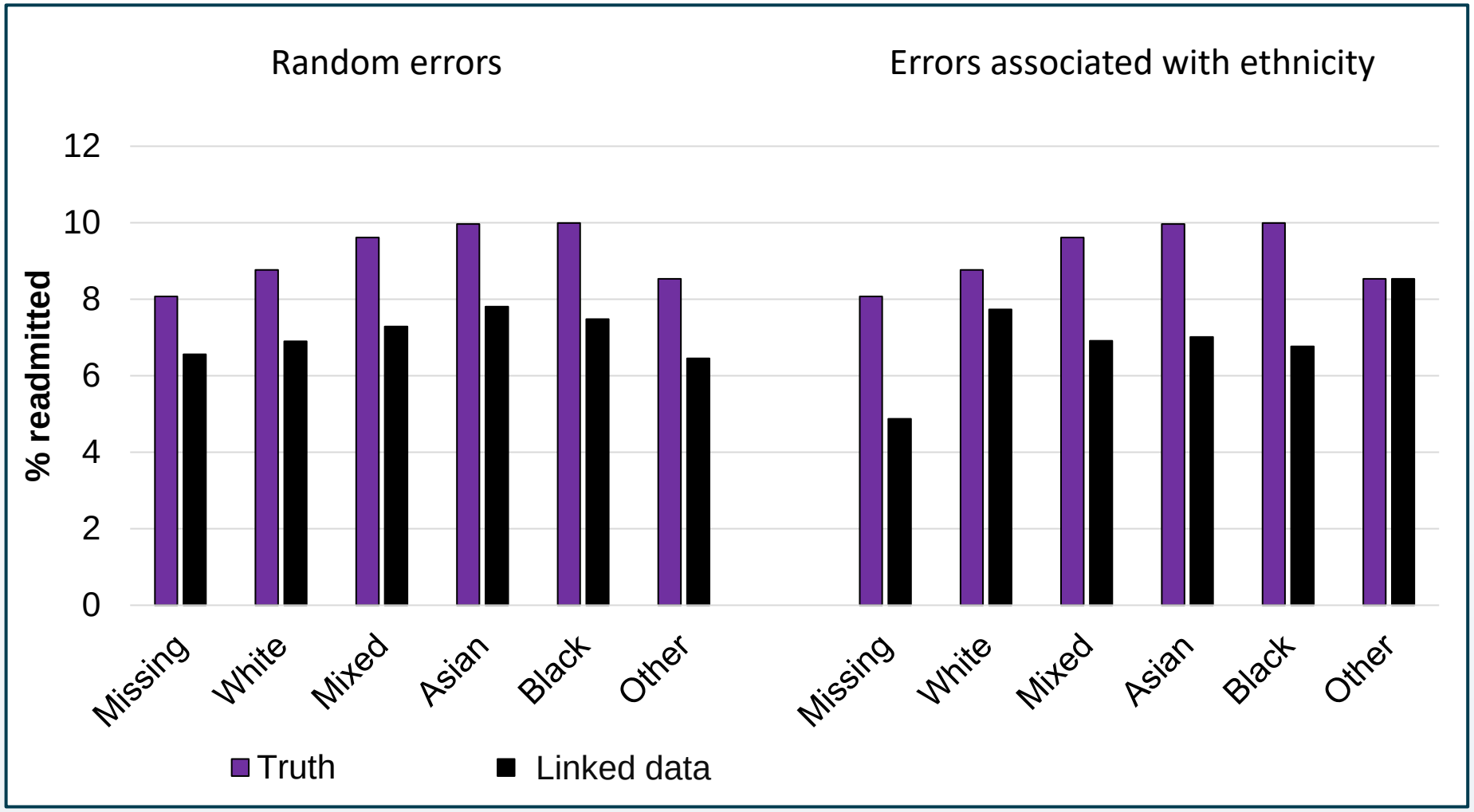
Doidge J, Harron K. Linkage error bias. Int J Epidemiol. 2019;dyz203.



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

	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



Harron, K., Hagger-Johnson, G., Gilbert, R. & Goldstein, H. Utilising identifier error variation in linkage of large administrative data sources. BMC Med Res Methodol 17, 23, doi:10.1186/s12874-017-0306-8 (2017).

How does linkage error lead to bias?

2: False matches

Misclassification or measurement error

Erroneous inclusion/exclusion in an analysis

‘Merging’ of multiple people’s records into one

Doidge J, Harron K. Linkage error bias. Int J Epidemiol. 2019;dyz203.

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 *(J Aging Health 2011)*

Solutions: Linkage quality assessment

Gold standard data

- Positive / negative controls
- Comparisons with external references in aggregate

Comparisons of linked / unlinked records

- Or of high / low quality records

Quality control checks

- Implausible scenarios

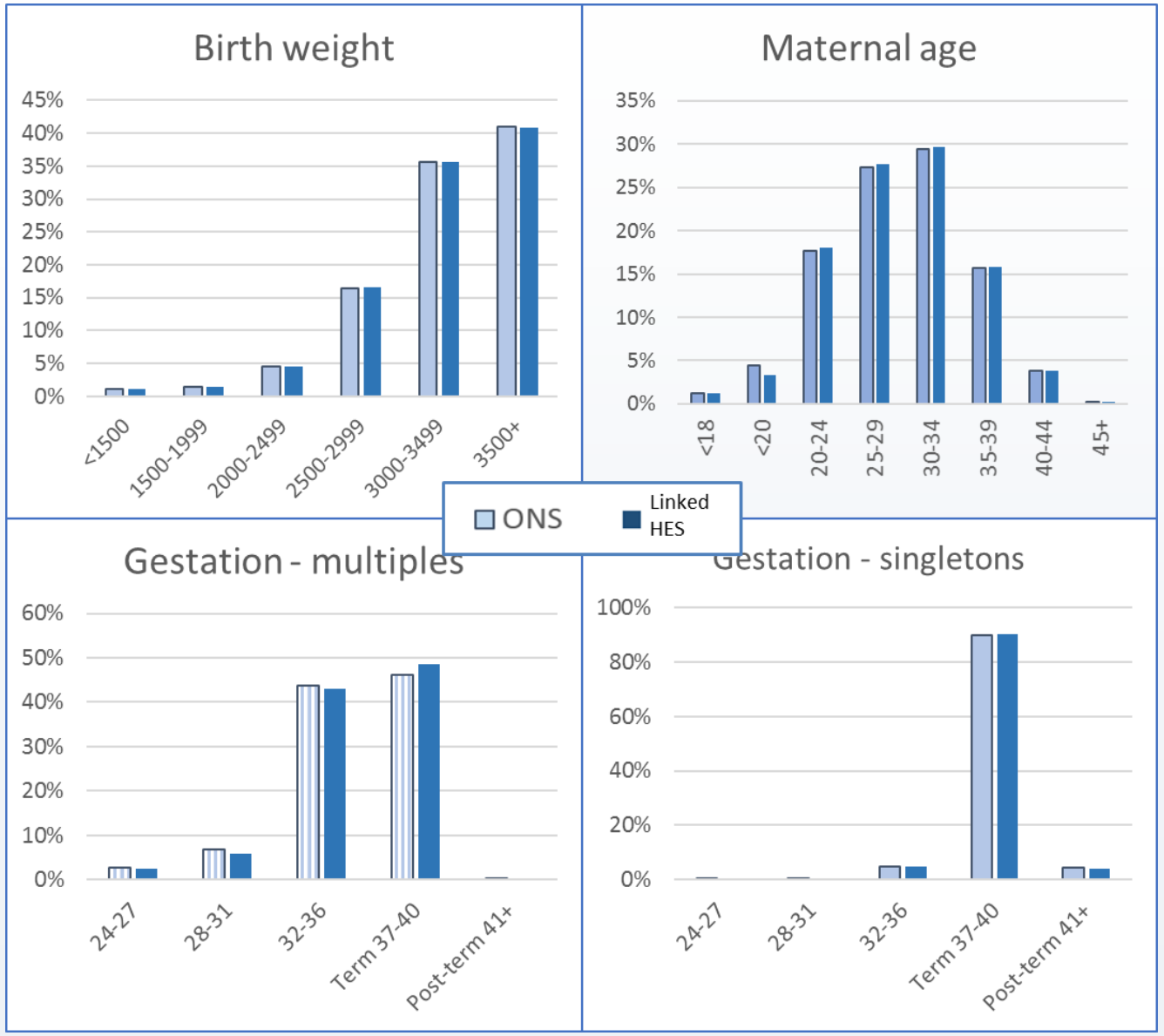
1. Harron KL, Doidge JC, Knight HE, et al. A guide to evaluating linkage quality for the analysis of linked data. International Journal of Epidemiology. 2017;46(5):1699-710.

Positive / negative controls

- Linking infection surveillance records with neonatal admission records - 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. *PloS One. 2019;14(12):e0226040-e*

- Linking pregnancies to birth registrations: pregnancies with abortive outcomes (-)
  - Paixão ES et al. 2019. Validating linkage of multiple population-based administrative databases in Brazil. *PloS One. 14(3):e0214050-e0214050*

Comparisons with external reference data



Harron K et al. 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

High / low quality records

		NHS Number				
		Available and valid		Not available or invalid		
	All	N	%	N	%	p-value†
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

Aldridge RW at al. 2015. Accuracy of Probabilistic Linkage Using the Enhanced Matching System for Public Health and Epidemiological Studies. PLoS ONE. 10(8):e0136179.

Quality control checks

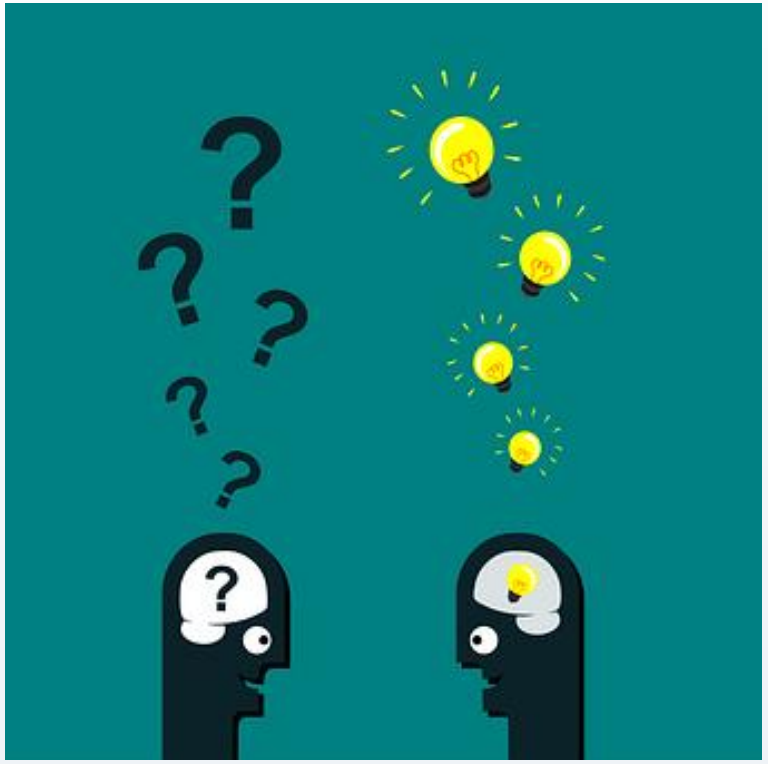
- Use evidence that two records do not belong to the same person to identify false-matches

  - E.g.,
  - Simultaneous admission in different part of the country
  - Admission following death
  - Linkage of prostate cancer records with female hospital records

	Infants (n = 733,770)		
	Not (n = 773,446)	Simultaneous Admission (N = 324)	p
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

Hagger-Johnson et al (2014) Identifying possible false matches in anonymized hospital administrative data without patient identifiers. Health Serv Res DOI:10.1111/1475-6773.12272

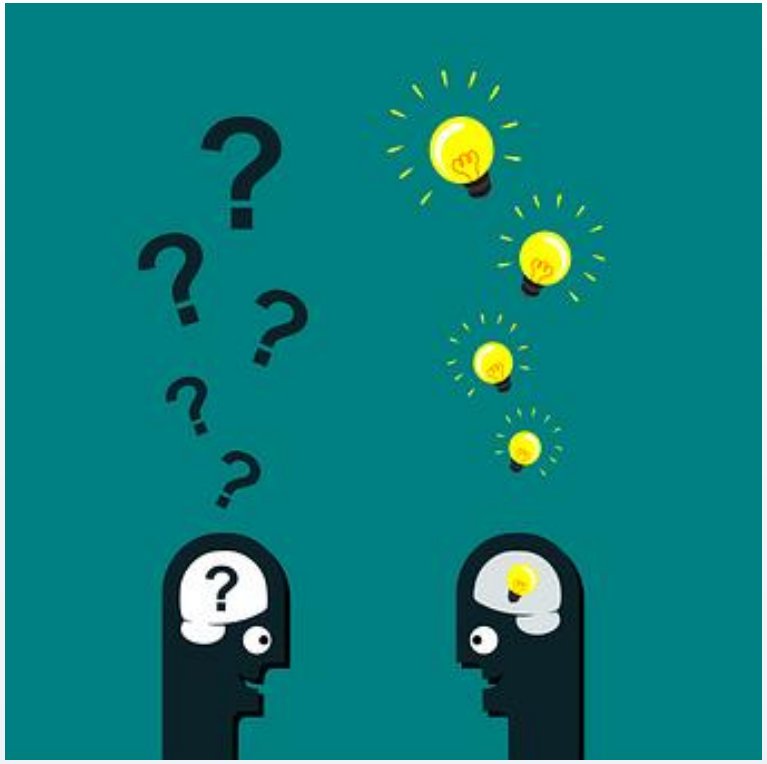
Solutions: handling linkage error in analyses



Treat as a missing data probelm

- *Uses information about uncertain links and record-level match quality*

Solutions: handling linkage error in analyses



Quantitative bias analysis

- *Uses group-level measures of linkage accuracy*
  - **Rates** of missed matches and false matches for different subgroups

Doidge et al. Prevalence of Down's Syndrome in England, 1998–2013: Comparison of linked surveillance data and electronic health Records. Int J Pop Data Sci 2020; in press

Nitsch D 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.

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
  - Imputation-based approaches
  - Quantitative bias analysis

Acknowledgements

Ruth Gilbert, Jan van der Meulen, James Doidge, Harvey Goldstein, Max Verfuerden, Ania Zylbersztejn, Hannah Knight, Nicolas Libuy, Ruth Blackburn

Funding:

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

CHIG
child health
informatics group

HDRUK
Health Data Research UK