Great Ormond Street Institute of Child Health

UCL

Transparency and reproducibility for linked administrative datasets

Katie Harron

UCL Great Ormond Street Institute of Child Health

February 2020

k.harron@ucl.ac.uk

CHIG
child health
informatics group

HDRUK
Health Data Research UK

- Why is reproducibility in data linkage important?
- What do we need to record and why?

Challenges

Quality of available identifiers

- Administrative data not designed for linkage
- Unique identifiers may not be present in all sources
- Choice of linkage methods

Linkage errors

- False matches and missed matches
- Can lead to substantially biased results
- Analysis needs to take uncertainty into account

Linkage methods

Deterministic (rule-based)

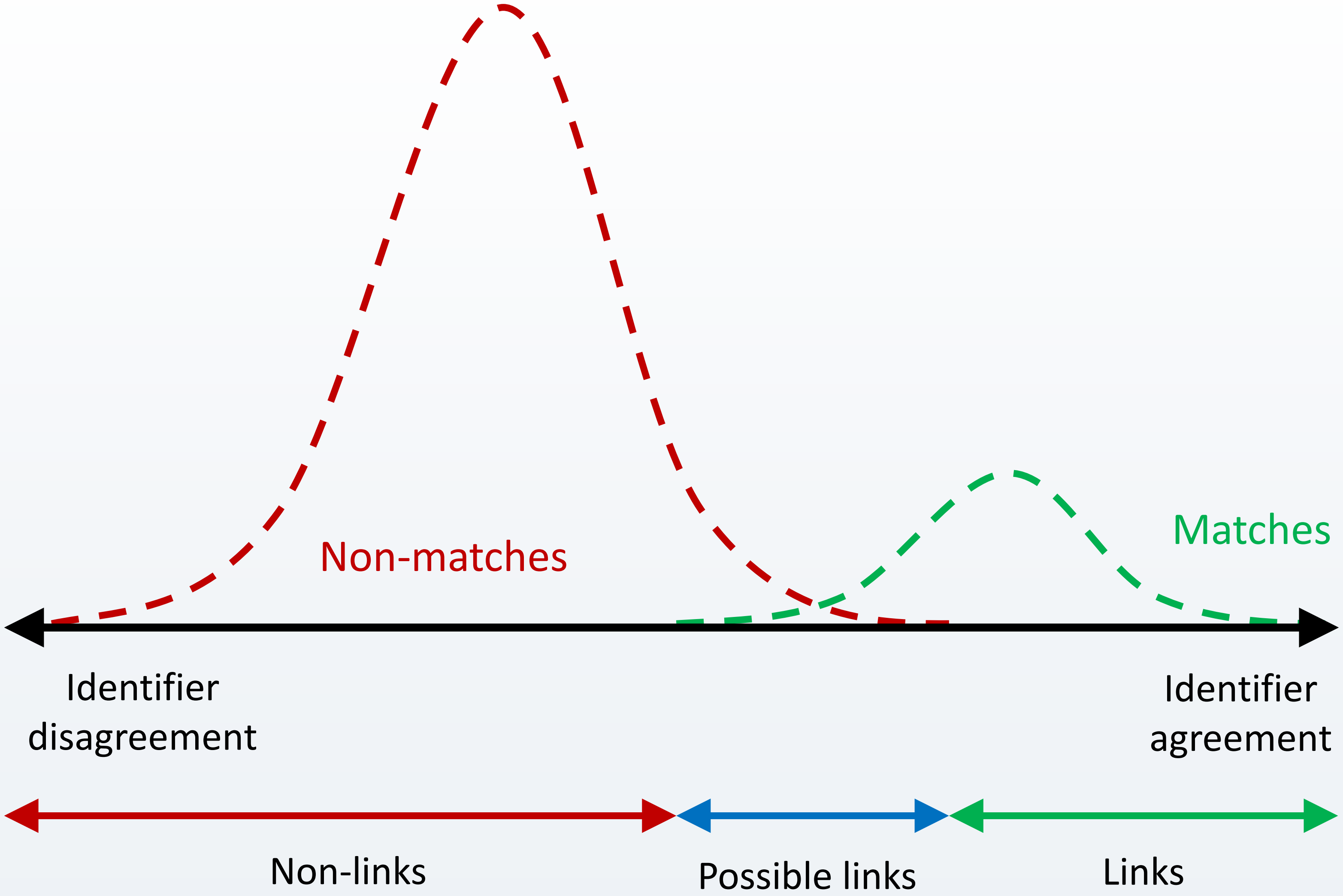
1
  - NHS Number
  - Sex
  - Date of Birth

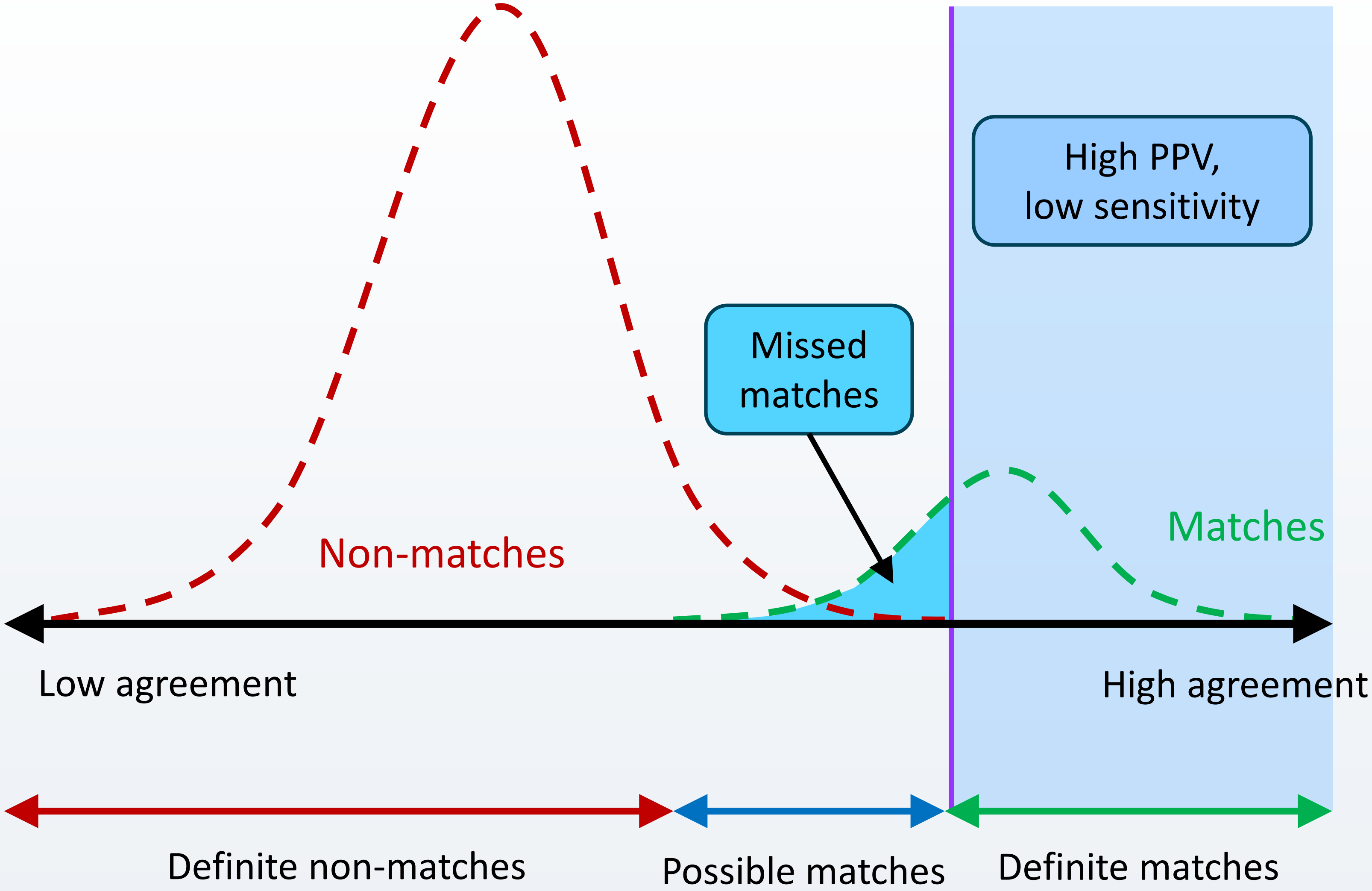
2
  - Hospital number
  - Postcode
  - Sex
  - Date of Birth

3
  - Postcode
  - Sex
  - Date of Birth

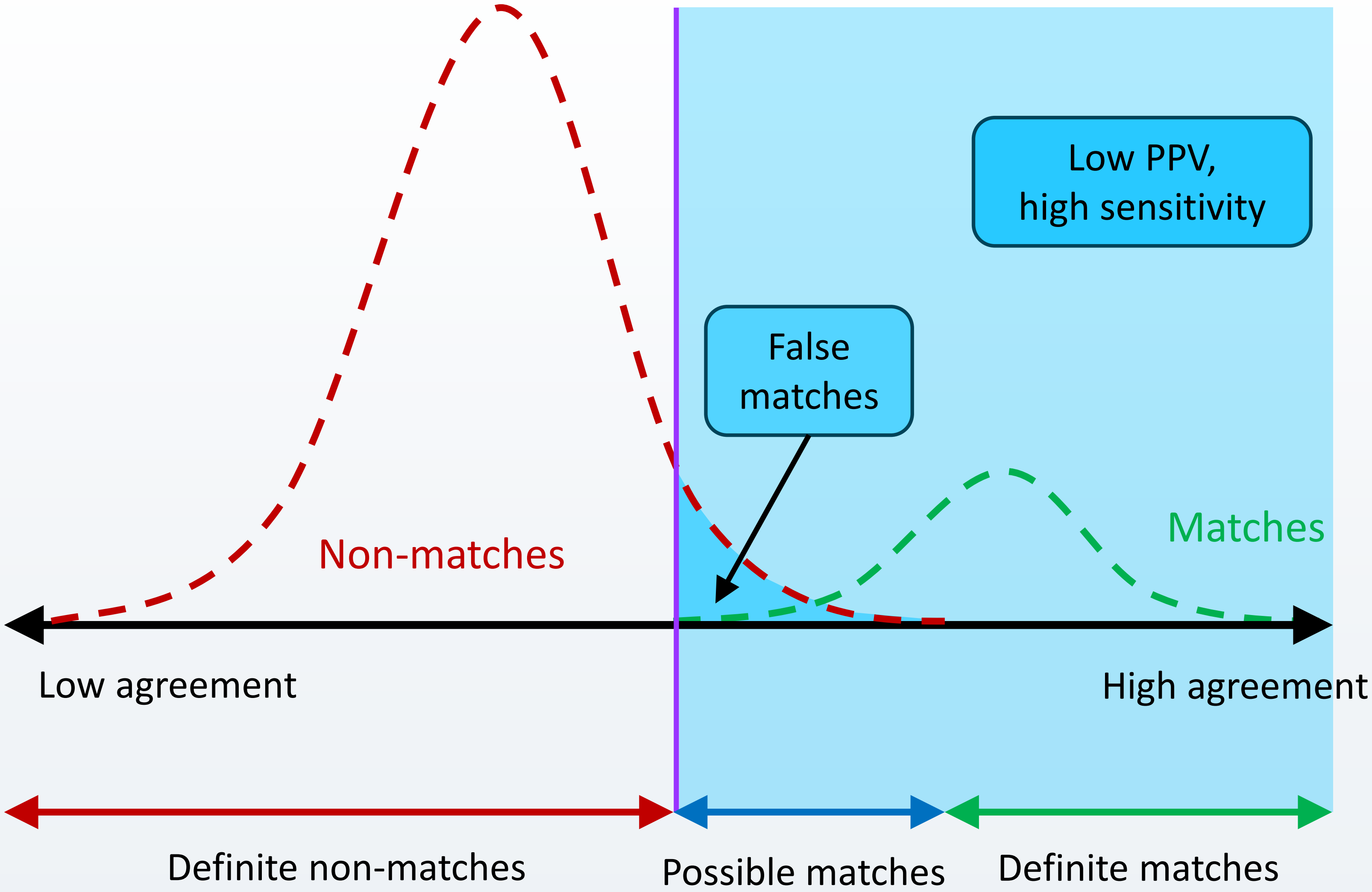
Probabilistic (score-based)

- Assigns a match weight representing the likelihood that two records belong to the same individual
- Takes into account how accurate and discriminative each identifier is

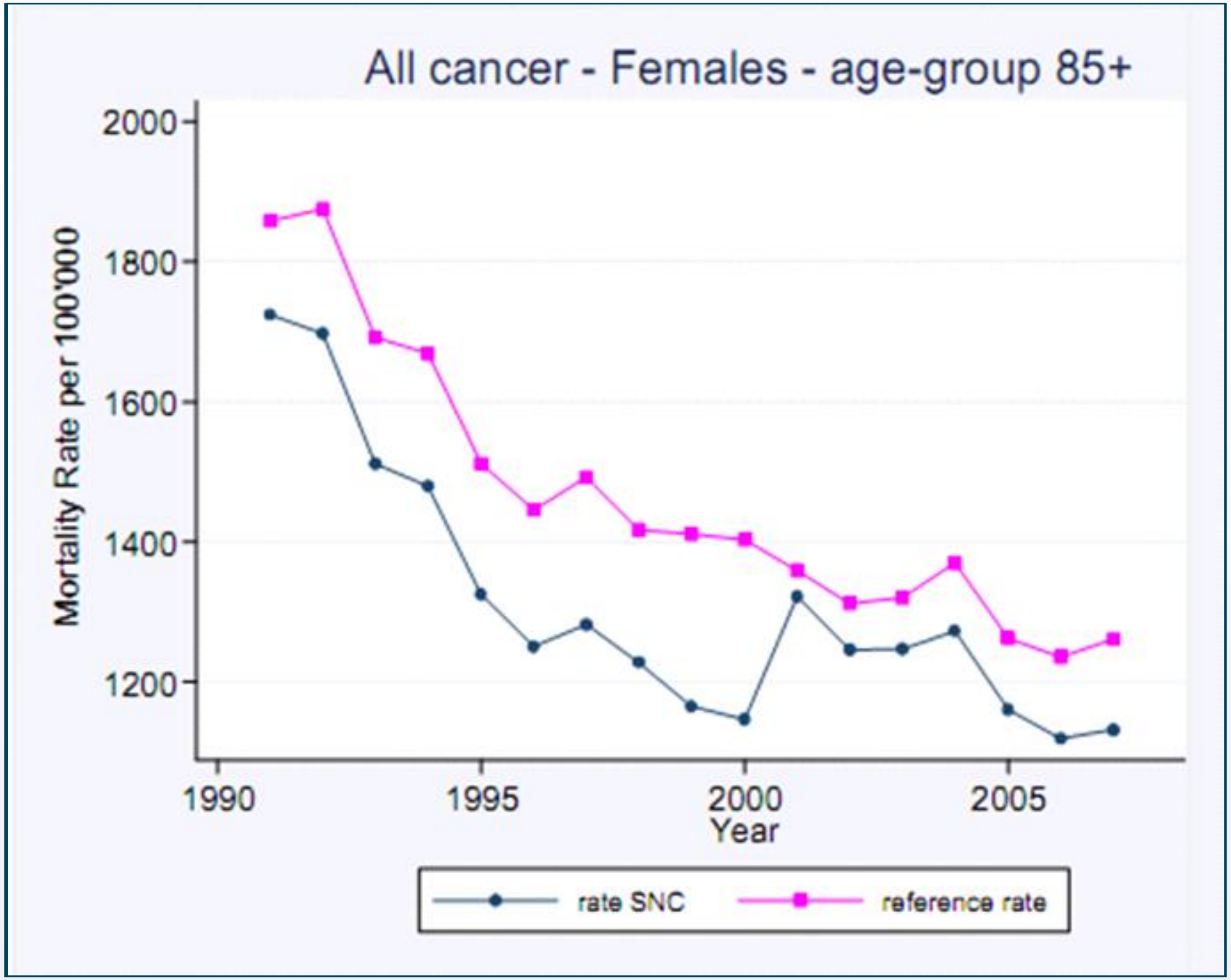

Non-matches
Matches
Identifier disagreement
Identifier agreement
Non-links
Possible links
Links



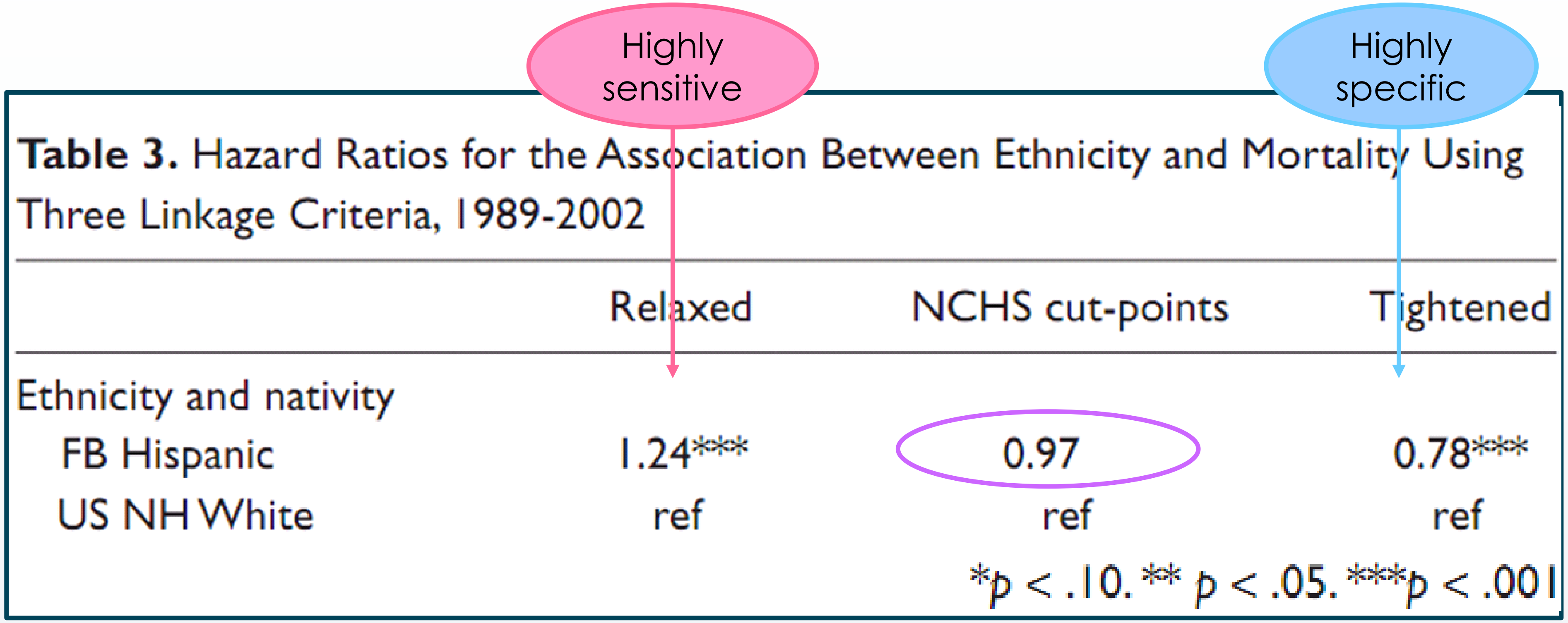
High PPV,
low sensitivity
Missed
matches
Matches
Non-matches
Low agreement
High agreement
Definite non-matches
Possible matches
Definite matches



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



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 *(J Aging Health 2011)*

What information do we need to record?

Details of the linkage algorithm

- How many linked at each stage?
- Were there any differences by subgroup?

Quality of identifiers

- Were there records that could never have been linked?

Quality assurance

- Estimates of rates of false / missed matches

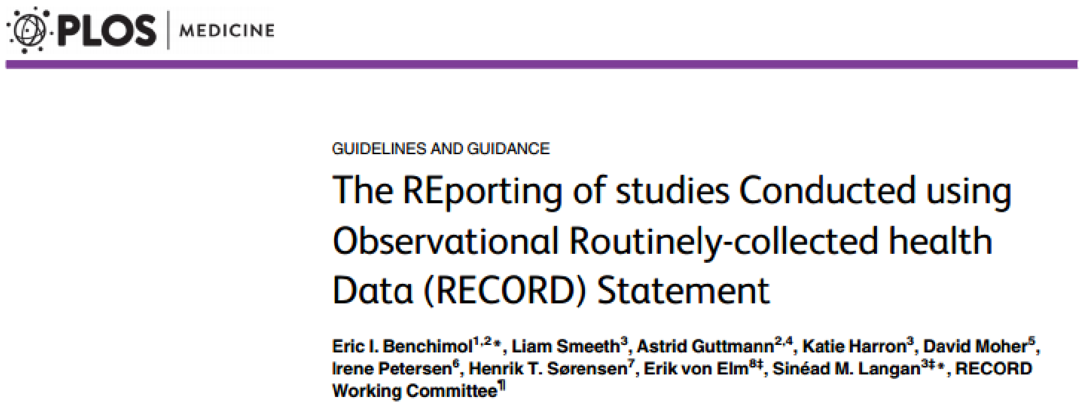
- Harron K, et al. (2012). "Opening the black box of record linkage." J Epidemiol Commun H 66(12): 1198.
- Harron K, et al. (2017). "A guide to evaluating linkage quality for the analysis of linked data." Int J Epidemiol **46**(5): 1699-1710.
- Doidge J and Harron K (2019). "Linkage error bias." Int J Epidemiol dyz203.

Guidelines

GUILD guidance

- GUidelines for Information about Linked Data
- Recommends information that should be shared at each step in the data linkage pathway
- To improve the quality and reproducibility of research based on linked data
- To minimise potential biases due to data processing and linkage error

Gilbert R et al. GUILD: GUidance for Information about Linking Datasets. *J Public Health* 2017;1-8.



PLOS | MEDICINE

GUIDELINES AND GUIDANCE

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

Eric I. Benchimol[1,2]*, Liam Smeeth[3], Astrid Guttmann[2,4], Katie Harron[3], David Moher[5], Irene Petersen[6], Henrik T. Sørensen[7], Erik von Elm[8‡], Sinéad M. Langan[3‡]*, RECORD Working Committee[¶]

http://record-statement.org/

National Statistician’s Quality Review on Data Linkage (2020)

https://gss.civilservice.gov.uk/guidances/quality/#national-statistician-s-quality-reviews-nsqrs-

Summary

- Reproducibility is important because results can change depending on how linkage was conducted

- There are various methods for evaluating linkage quality and accounting for bias due to linkage within analysis
  - Communication between data linkers and data users is key
  - Guidelines are available

- Accounting for linkage error and uncertainty will lead to more robust research

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.