# Linking administrative data for research

Katie Harron, Sir Henry Wellcome Postdoctoral Fellow

Department of Health Services Research and Policy, LSHTM

February 2017

wellcome trust Fellow

# A statistical definition

**"a merging that brings together information from two or more sources of data with the object of consolidating facts concerning an individual or an event that are not available in any separate record"**

Organisation for Economic Co-operation and Development (OECD)
Glossary of Statistical Terms

# Record linkage for health data

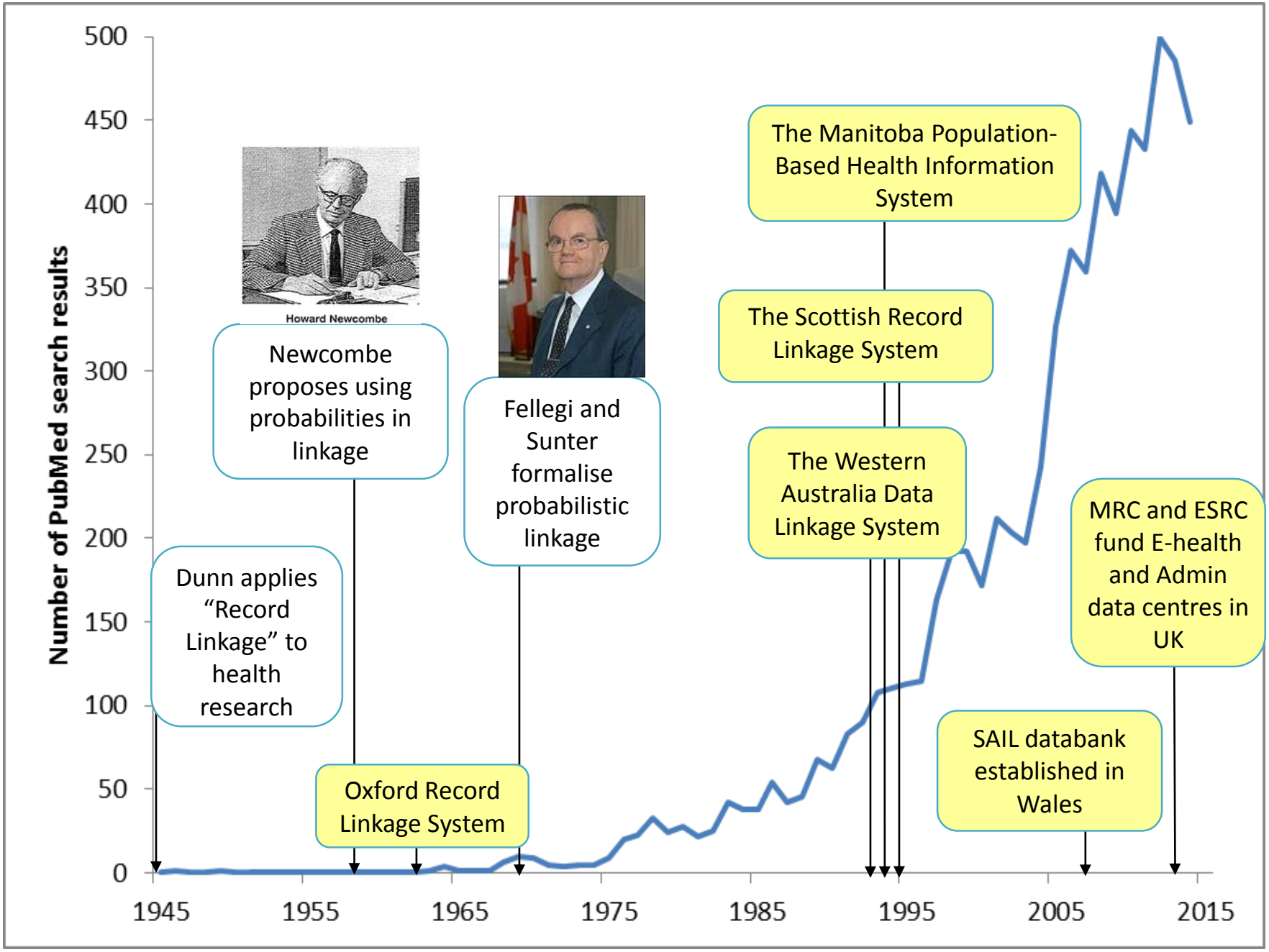
Each person in the world creates a Book of Life.

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

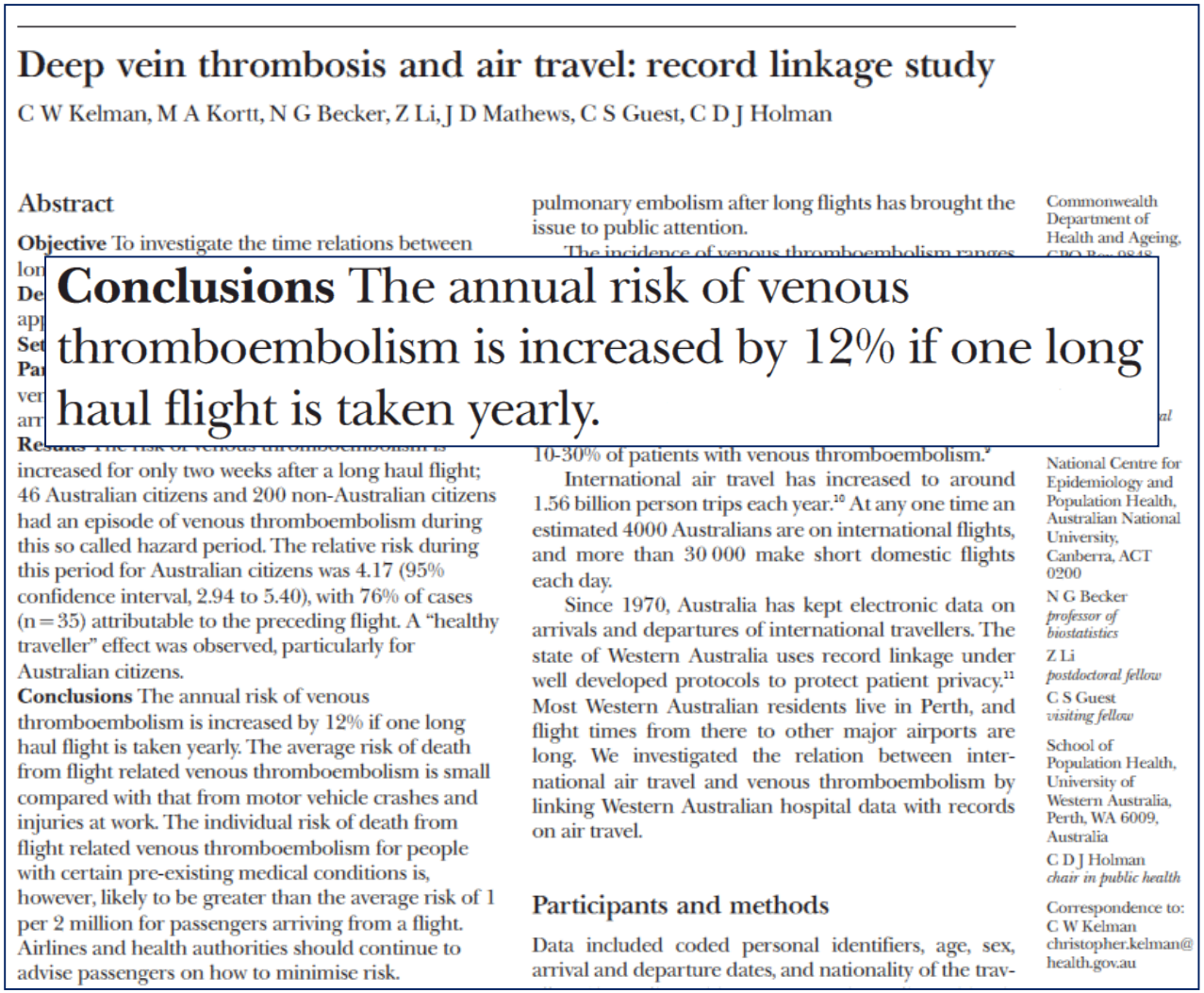

Number of PubMed search results
500
450
400
350
300
250
200
150
100
50
0
1945
1955
1965
1975
1985
1995
2005
2015
Howard Newcombe
Newcombe proposes using probabilities in linkage
Dunn applies "Record Linkage" to health research
Oxford Record Linkage System
Fellegi and Sunter formalise probabilistic linkage
The Manitoba Population-Based Health Information System
The Scottish Record Linkage System
The Western Australia Data Linkage System
MRC and ESRC fund E-health and Admin data centres in UK
SAIL databank established in Wales

# Opportunities and challenges linking administrative data

**Opportunities**

+ population-level resource
+ (potentially) lower risk of selection bias - generalisabilty
+ allows evaluation of rare events / hard to reach subgroups
+ detailed longitudinal trajectories
+ cost effective – exploits existing data
+ 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**
app
**Set**
**Pa**
ver
arr
**Re** The risk of venous thromboembolism is 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, GPO Box 9848

al

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*

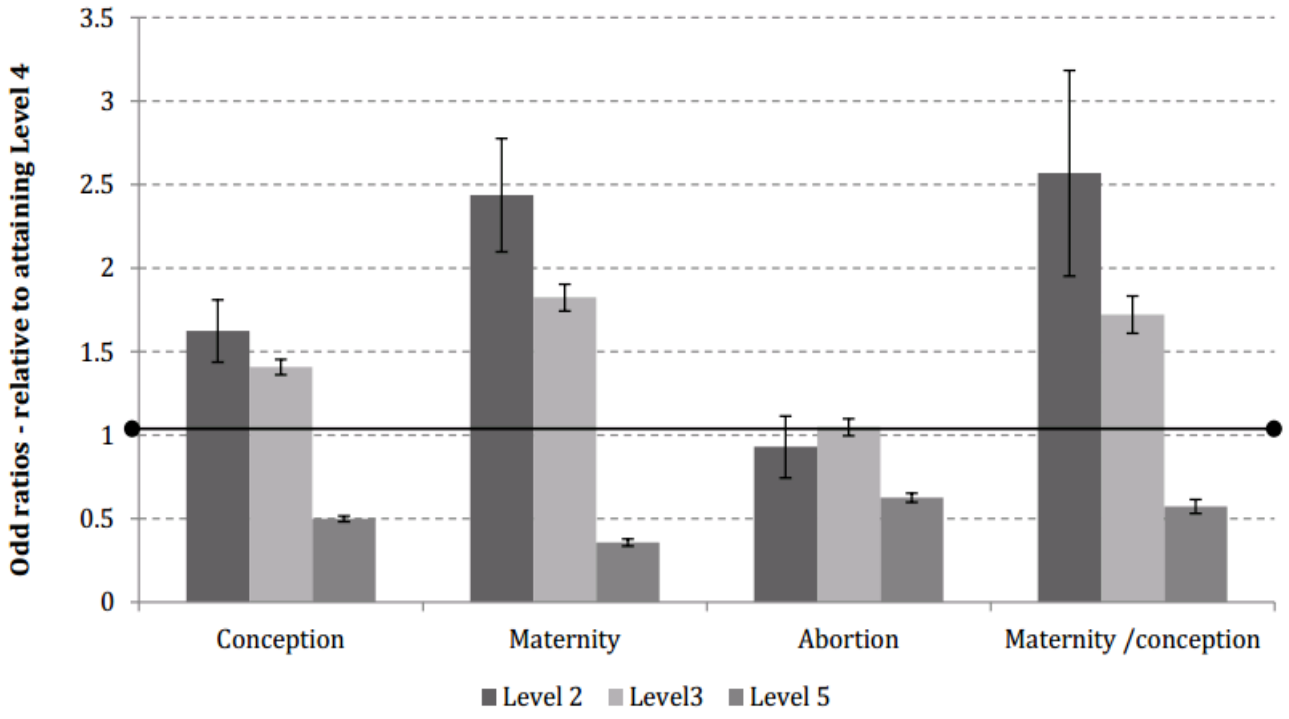
Correspondence to:
C W Kelman
christopher.kelman@health.gov.au


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 linking administrative data

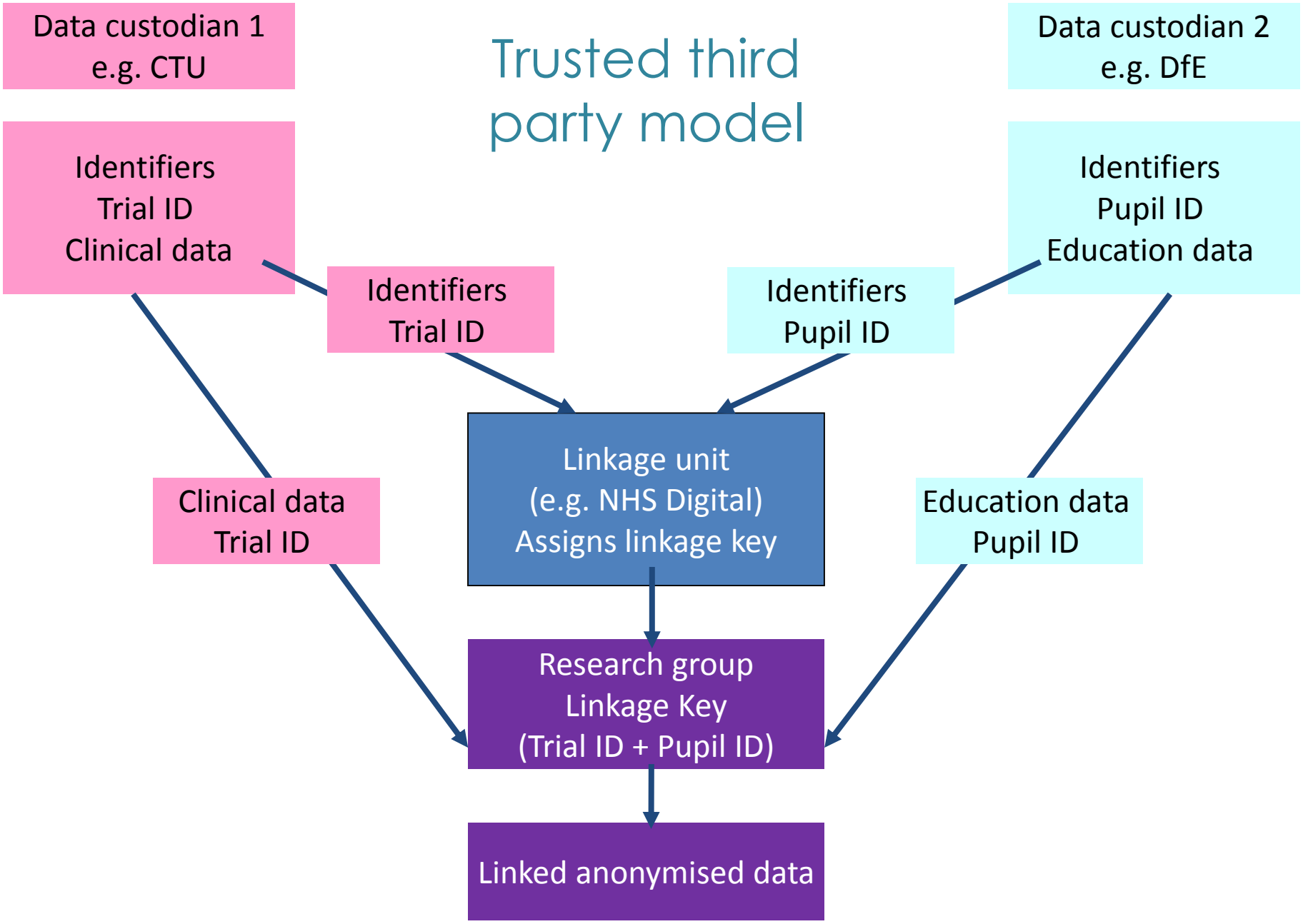
**Opportunities**

+ population-level resource
+ (potentially) lower risk of selection bias - generalisabilty
+ detailed longitudinal healthcare trajectories
+ allows evaluation of rare events / hard to reach subgroups
+ cost effective – exploits existing data
+ answer novel research questions

**Challenges**

- uncertainty about data quality
- lack of unique identifiers for linkage
- data security considerations

# Trusted third party model



# Linking hospital records for mothers and babies

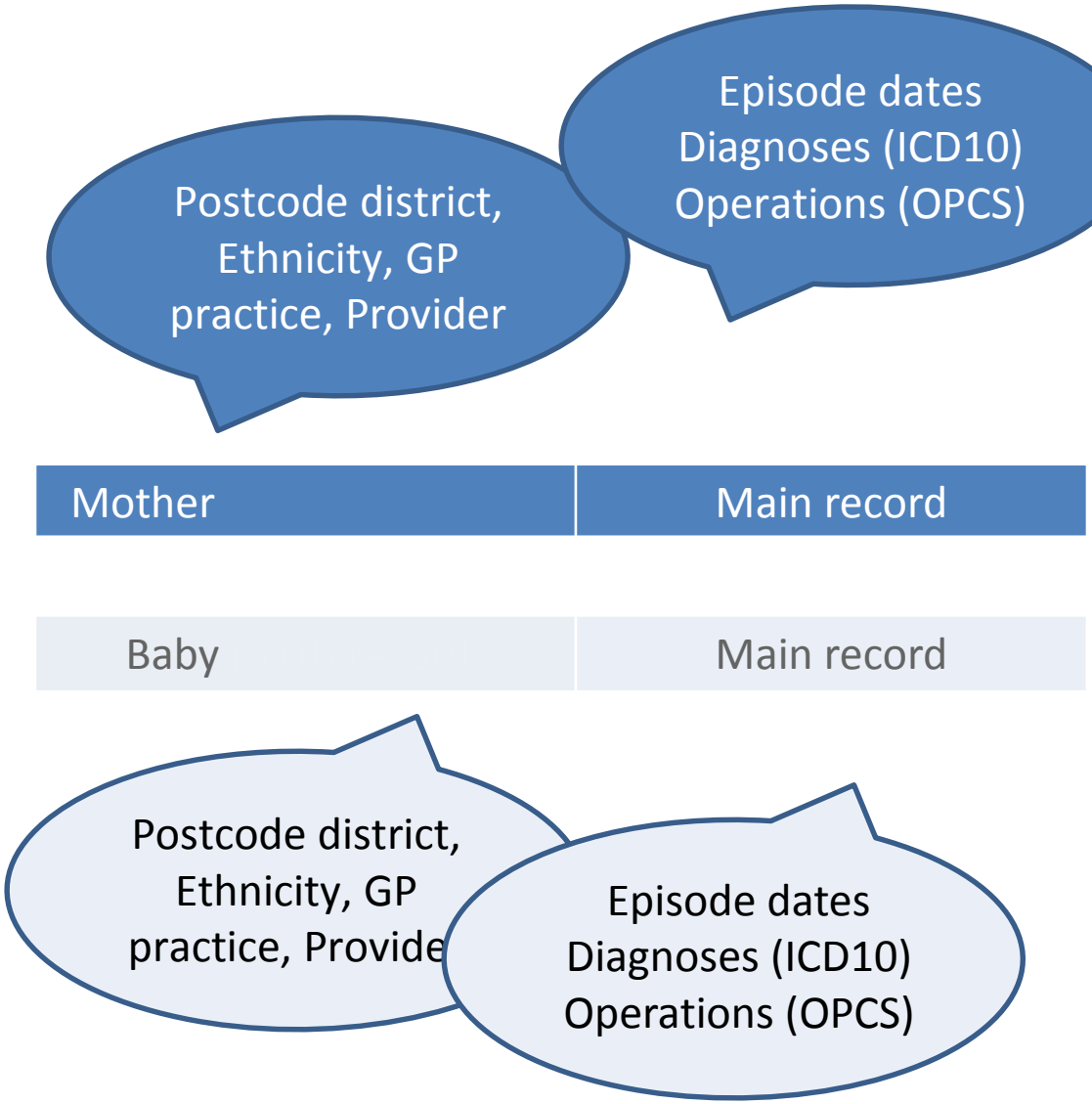
**Opportunities**

- novel research questions, involving rare outcomes
  - → induction of labour and perinatal mortality / neonatal morbidity
  - → maternal mortality following neonatal abstinence syndrome

**Challenges**

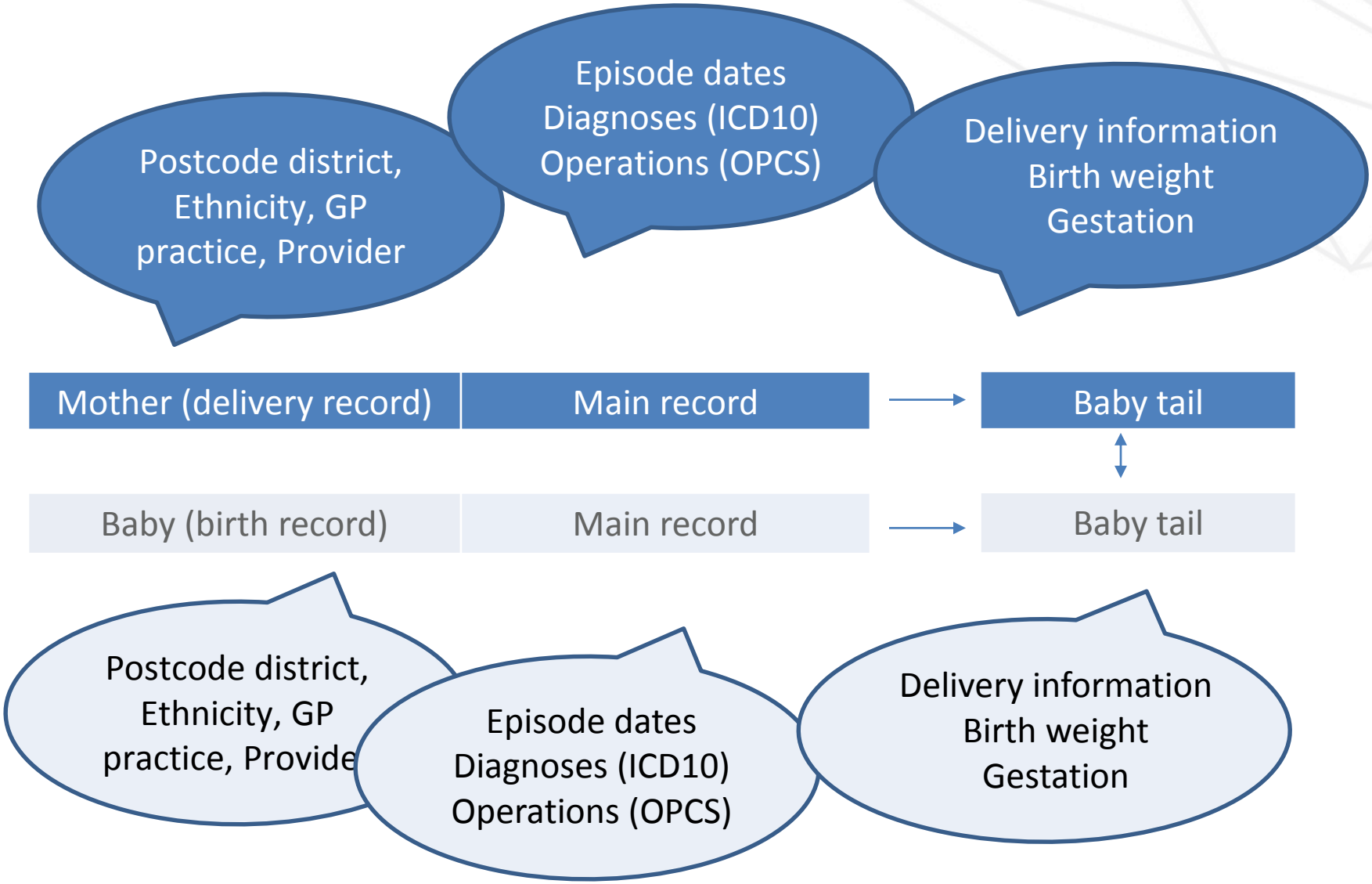
- uncertainty about data quality
  - → can we use linkage to better understand / improve data quality?
- lack of unique identifiers for linkage
  - → how can we handle bias due to linkage error?

# Hospital Episode Statistics




Herbert et al. Data Resource Profile: Hospital Episode Statistics Admitted Patient Care (HES APC). Int J Epidemiol in press (2017).

# Hospital Episode Statistics



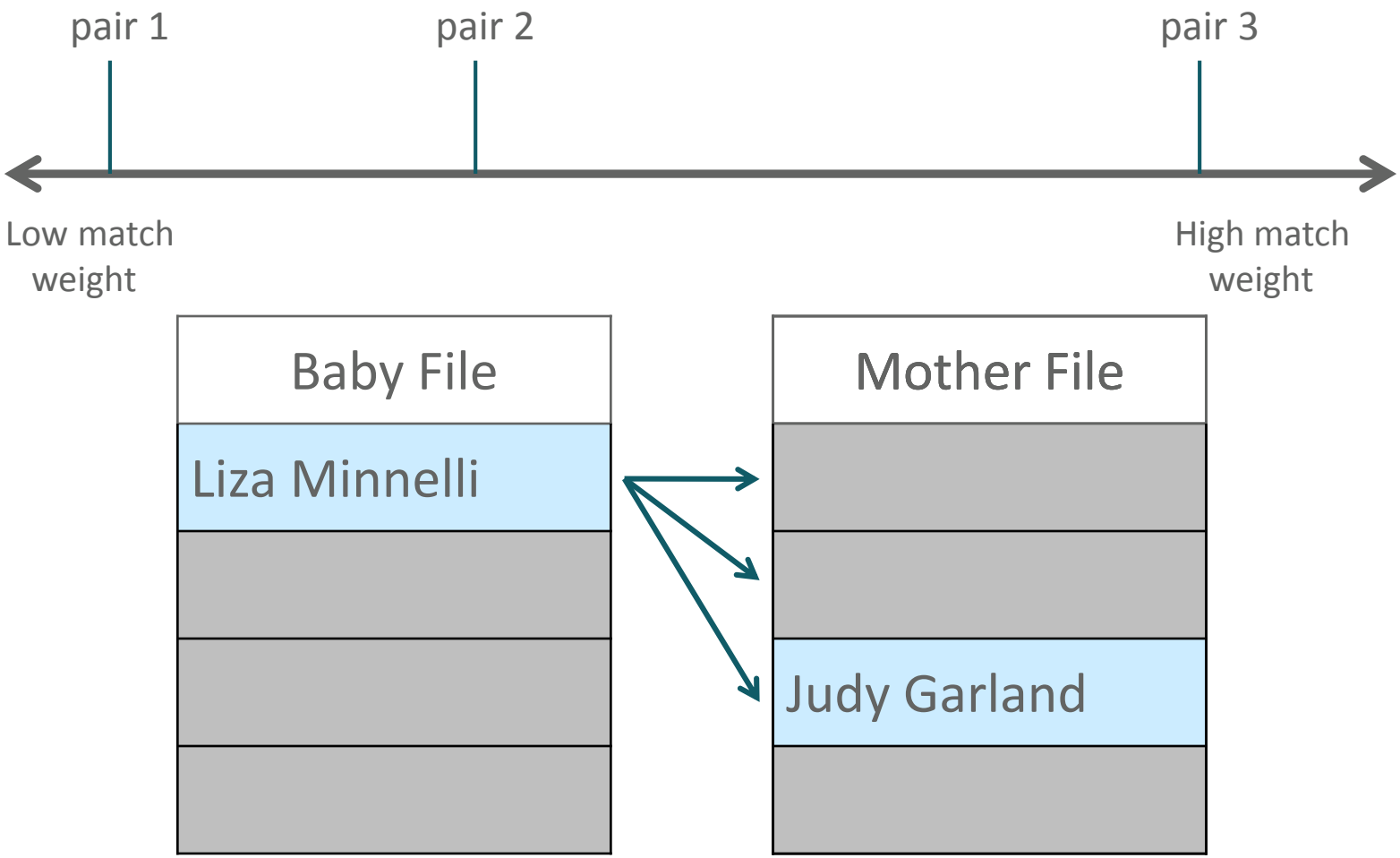

Herbert et al. Data Resource Profile: Hospital Episode Statistics Admitted Patient Care (HES APC). Int J Epidemiol in press (2017).

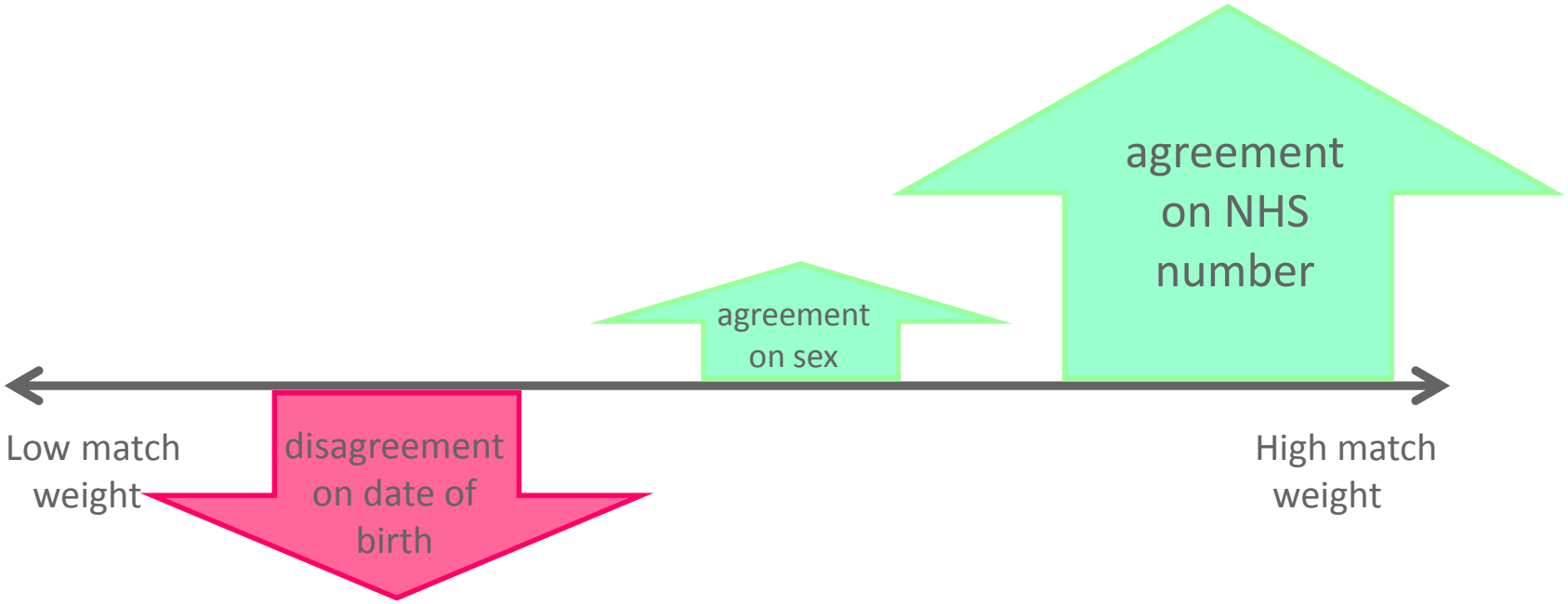
# Linkage methods

- Deterministic
  - Rule-based approach, often looking for exact agreement on a number of identifiers
- Probabilistic
  - Uses the conditional probability that identifiers on different records will agree
    - Given records belong to the same person
    - Given records belong to different people (~ agreement by chance)

Administrative Data Research Network: Introduction to Data Linkage. https://adrn.ac.uk/media/1372/guide1-data-linkage-katie-harron-aug-2016.pdf

# Probabilistic linkage






P(**γ**=1 | **M**) = ***m-probability***

the probability of agreement given the records from same subject

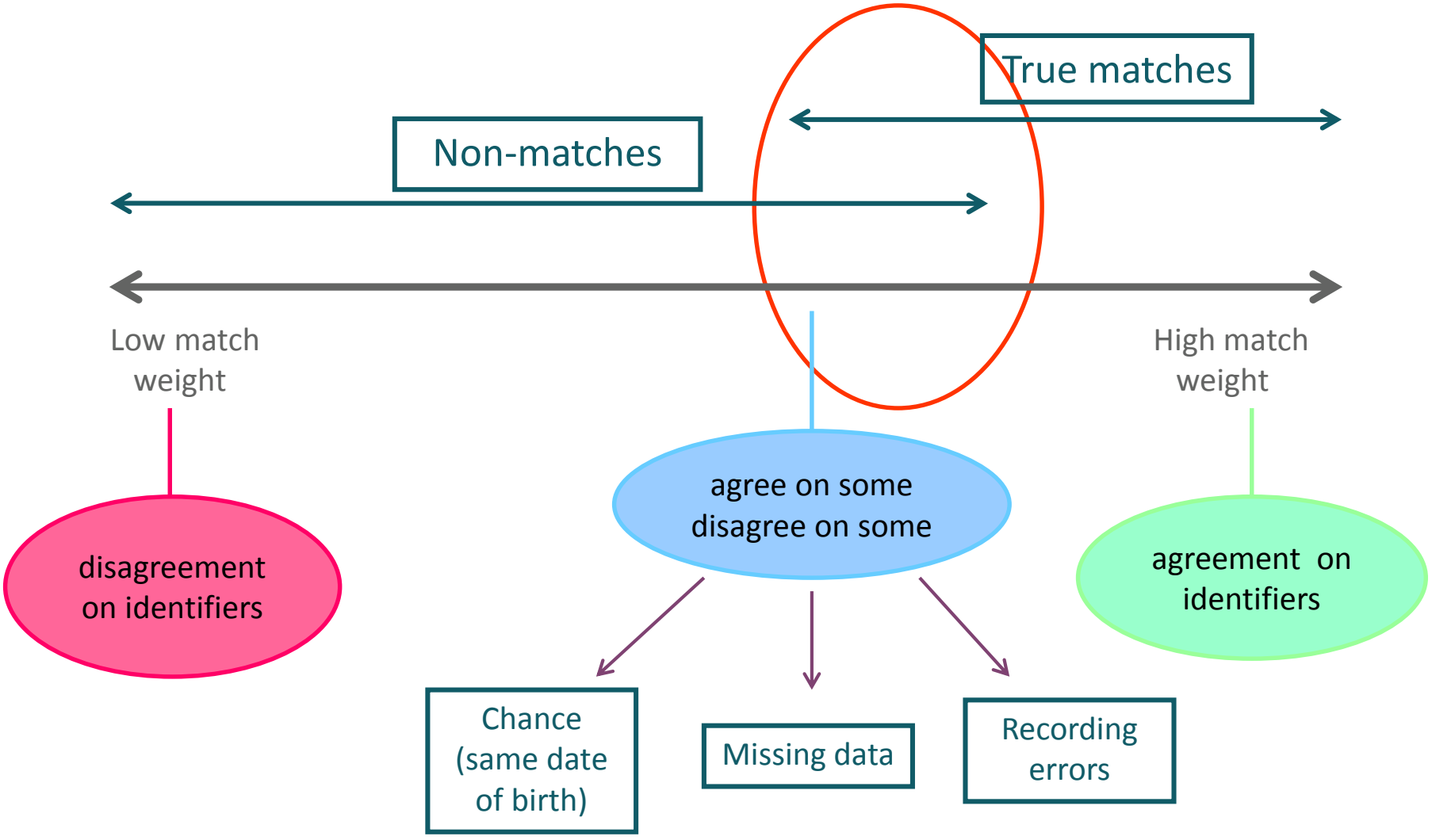
P(**γ**=1 | **U**) = ***u-probability*** =

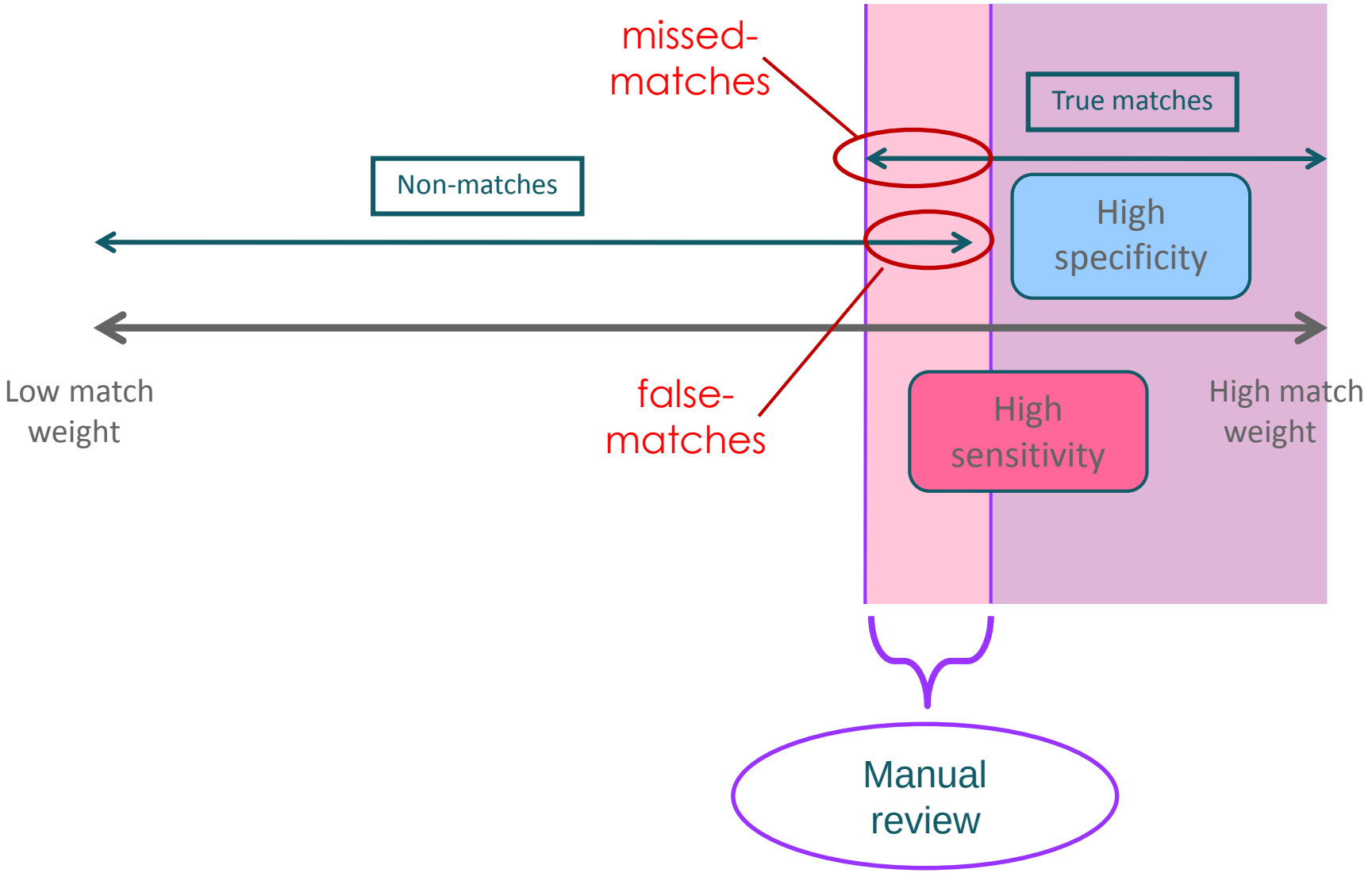
the probability of agreement given the records from different subjects

**Log ratio = w** = $\log_2 (m/u)$ if identifiers agree

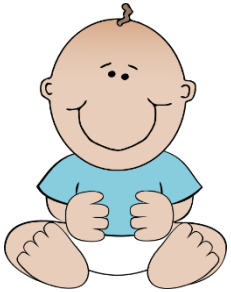
$\log_2 [(1-m)/(1-u)]$ if identifiers disagree

**Match weight = W** = $\sum w_i$


True matches
Non-matches
Low match weight
High match weight
disagreement on identifiers
agree on some disagree on some
agreement on identifiers
Chance (same date of birth)
Missing data
Recording errors


missed-
matches
True matches
Non-matches
High
specificity
Low match
weight
false-
matches
High
sensitivity
High match
weight
Manual
review

# Linkage

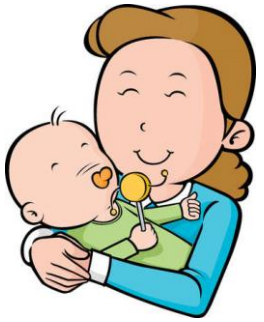


**Baby records 2012**
N = 673,055

**Maternal records 2012**
N=671,436

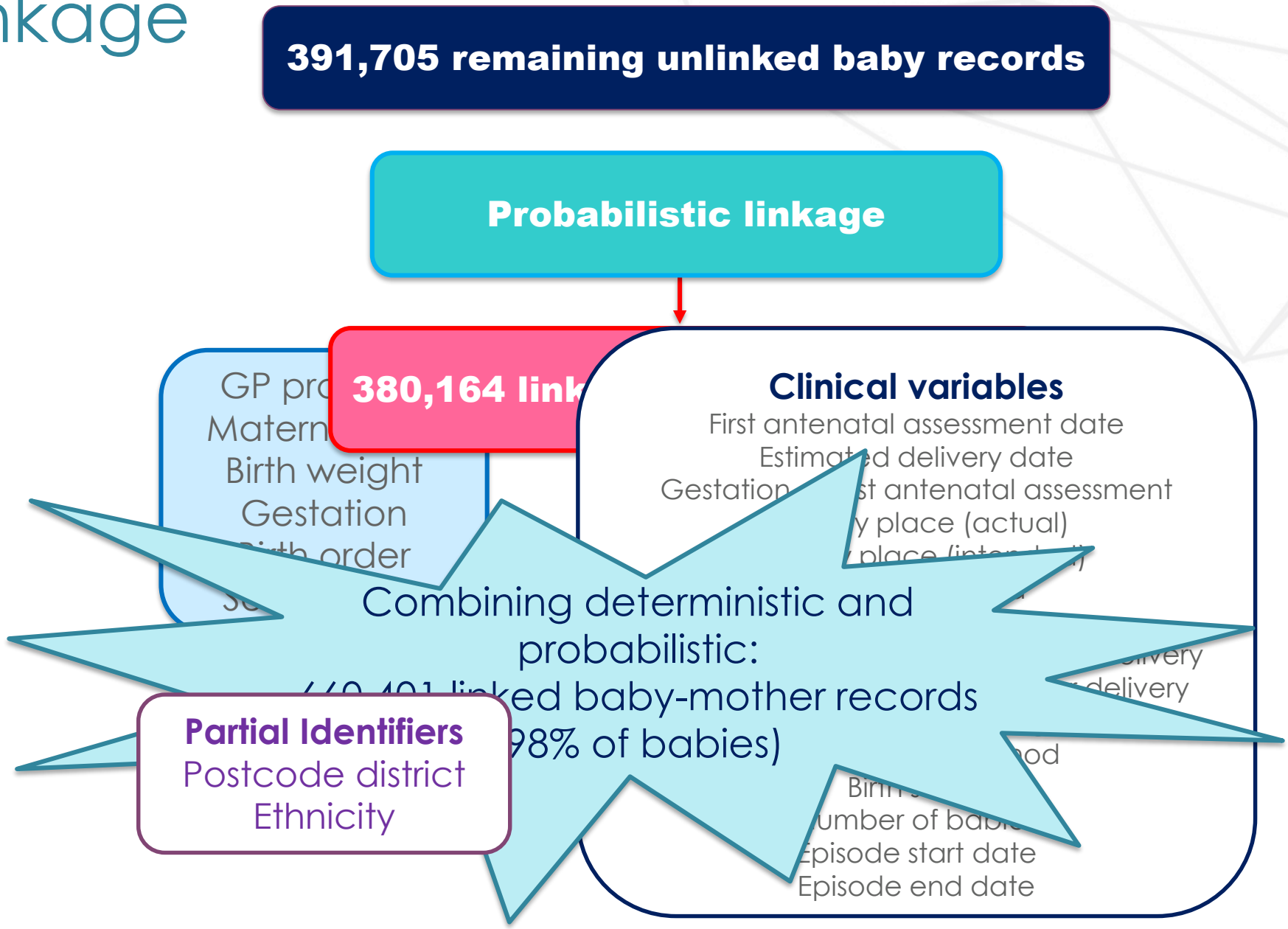
**Deterministic linkage:**
GP practice
Maternal age
Birth weight
Gestation
Birth order
Sex of baby

**280,939 linked baby records (42%)**



Harron et al. Linking data for mothers and babies in de-identified electronic health data. PLoS One. 2016;11(10):e0164667.

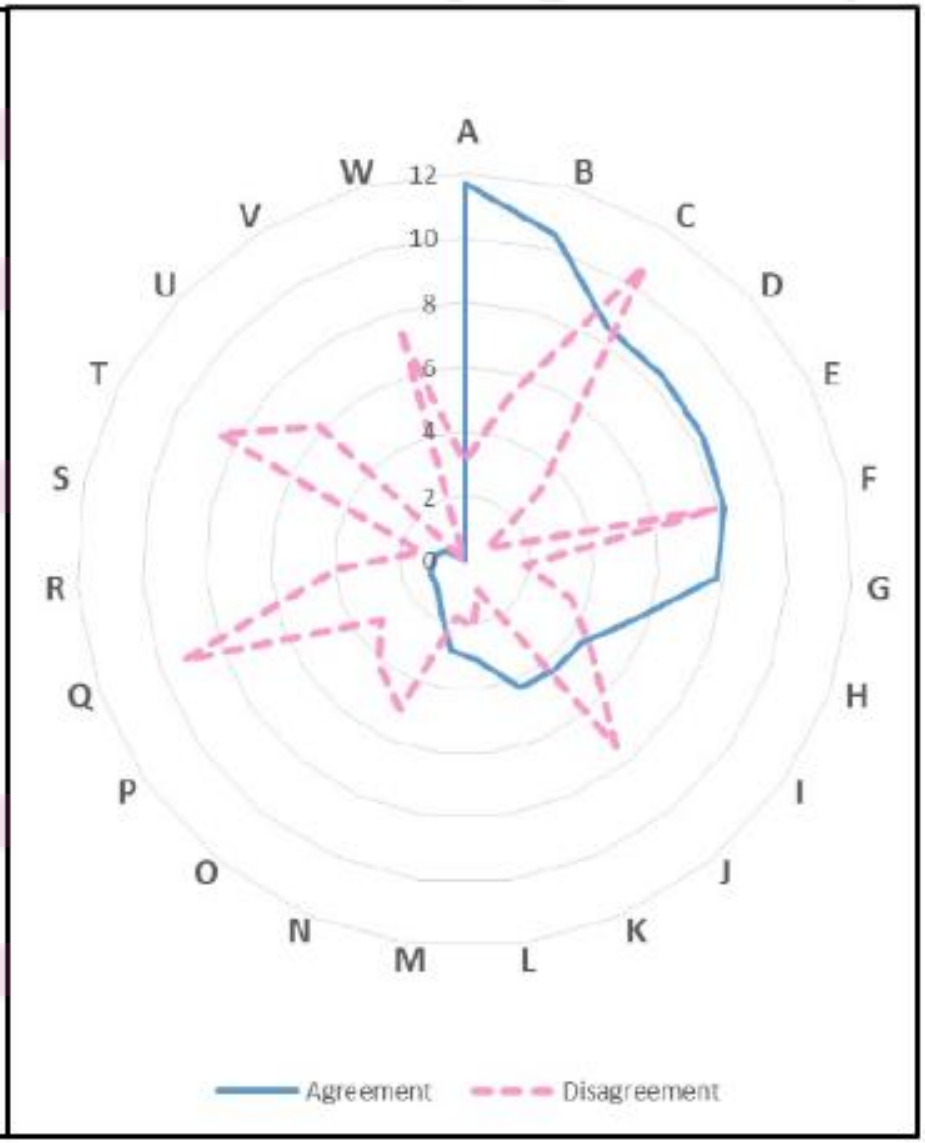
# Linkage




Harron et al. Linking data for mothers and babies in de-identified electronic health data. PLoS One. 2016;11(10):e0164667.

# Probabilistic match weights

| | |
|---|---|
| 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 |



# Understanding data quality

Gestation complete in 84%
→ Preterm birth rate = 6.3%


With linkage of information from baby record:

Completeness of gestation increases: from 84% → 92%

Preterm birth rate increases: from 6.3% → 6.7%

# Understanding data quality

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 |
|---|---|---|

# Understanding data quality

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 |
|---|---|---|

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

# Understanding data quality

With linkage of information from baby record:

800/1558 stillbirth conflicts resolved by triangulating information held on mother/baby records

- Checking ICD10 codes, birth status, length of stay
- 0.1% of records unresolved

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

# Lack of reliable unique identifiers → Linkage error

| | | **Match status** | |
|---|---|---|---|
| | | Match (same mother-baby pair) | Non-match (different mother-baby pair) |
| **Link status** | Link | Identified match | False match |
| | Non-link | Missed match | Identified non-match |

# The linkage problem

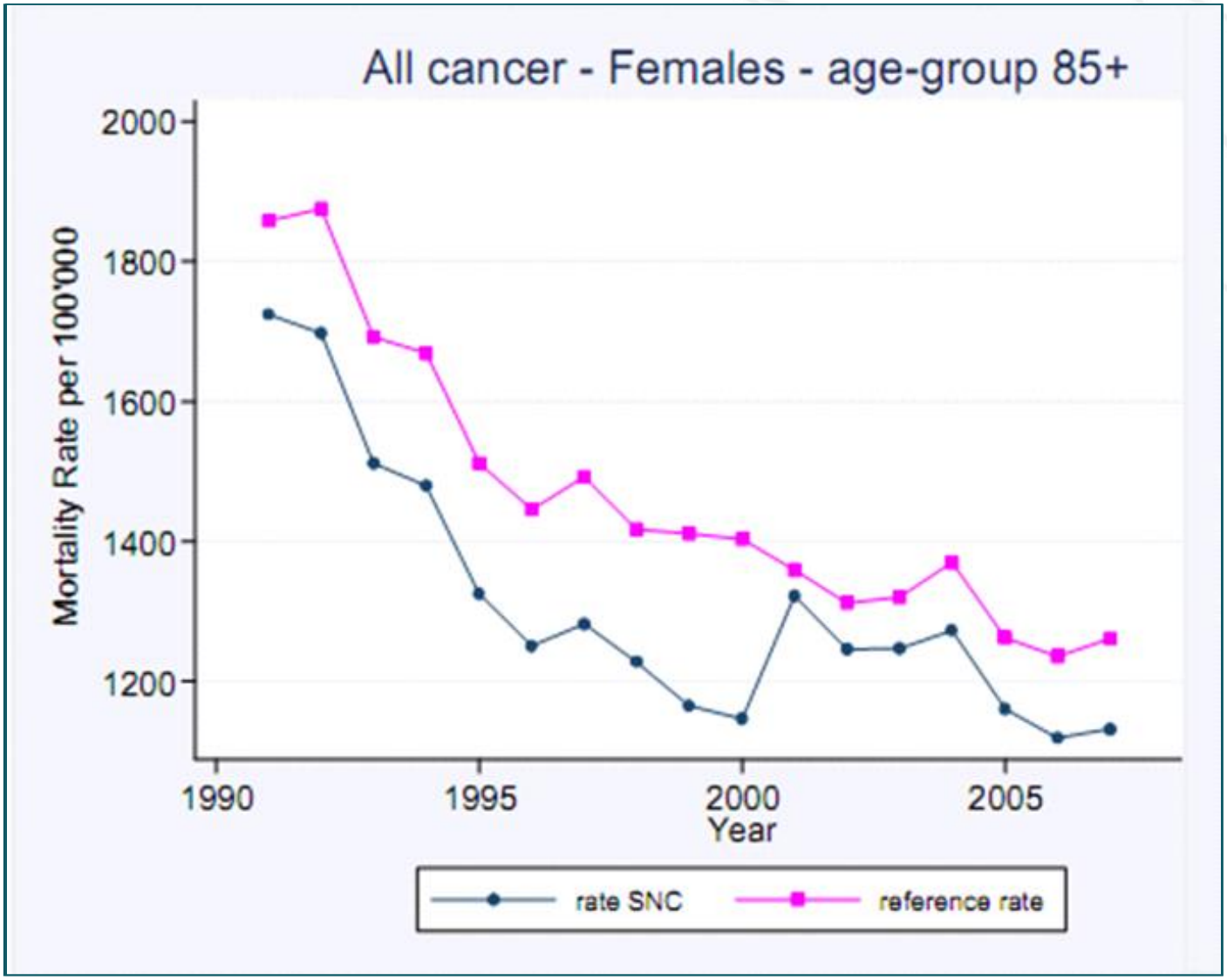
- Small amounts of linkage error can result in substantially biased results

- 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 |

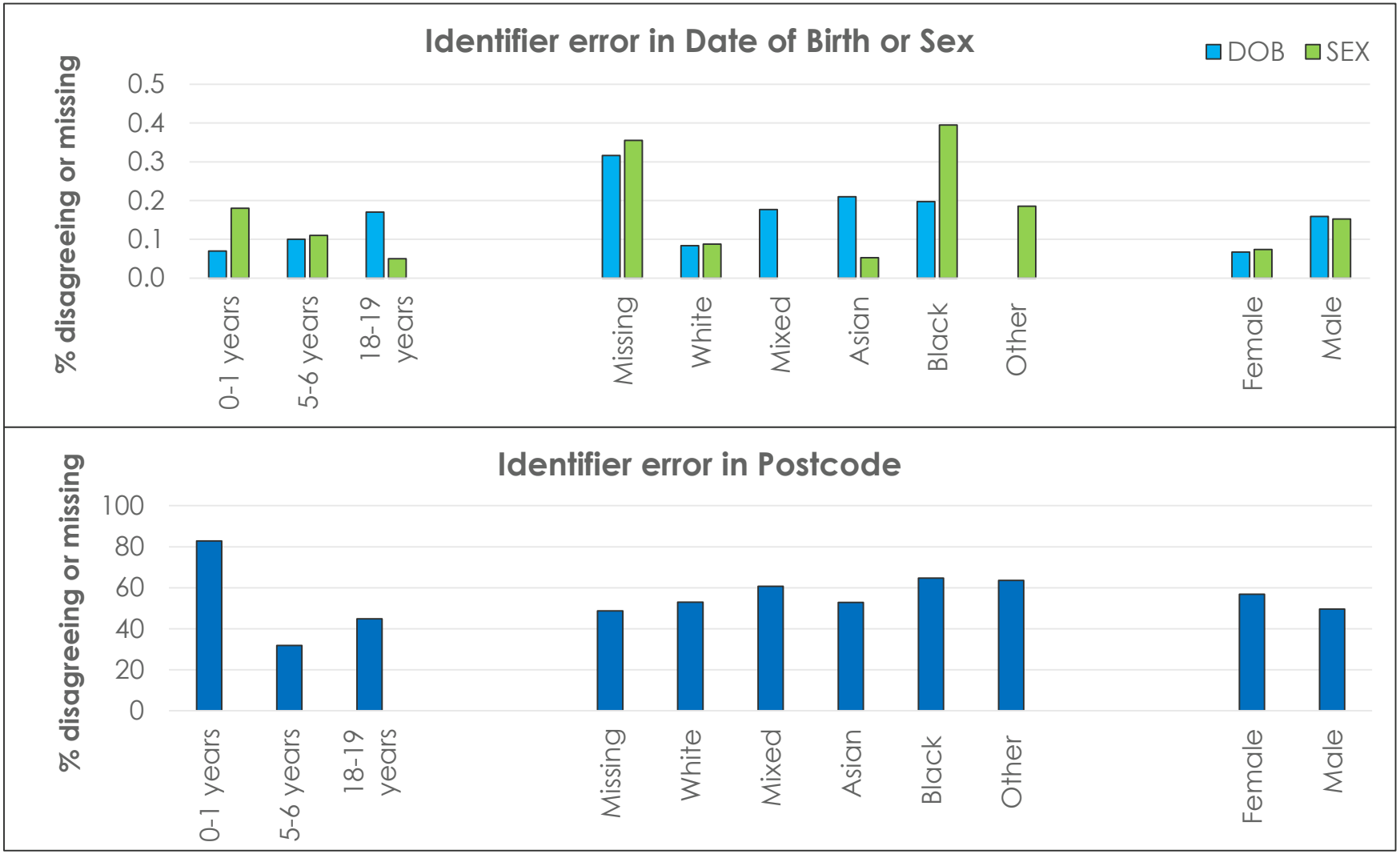
*$p < .10$. ** $p < .05$. ***$p < .001$

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 |

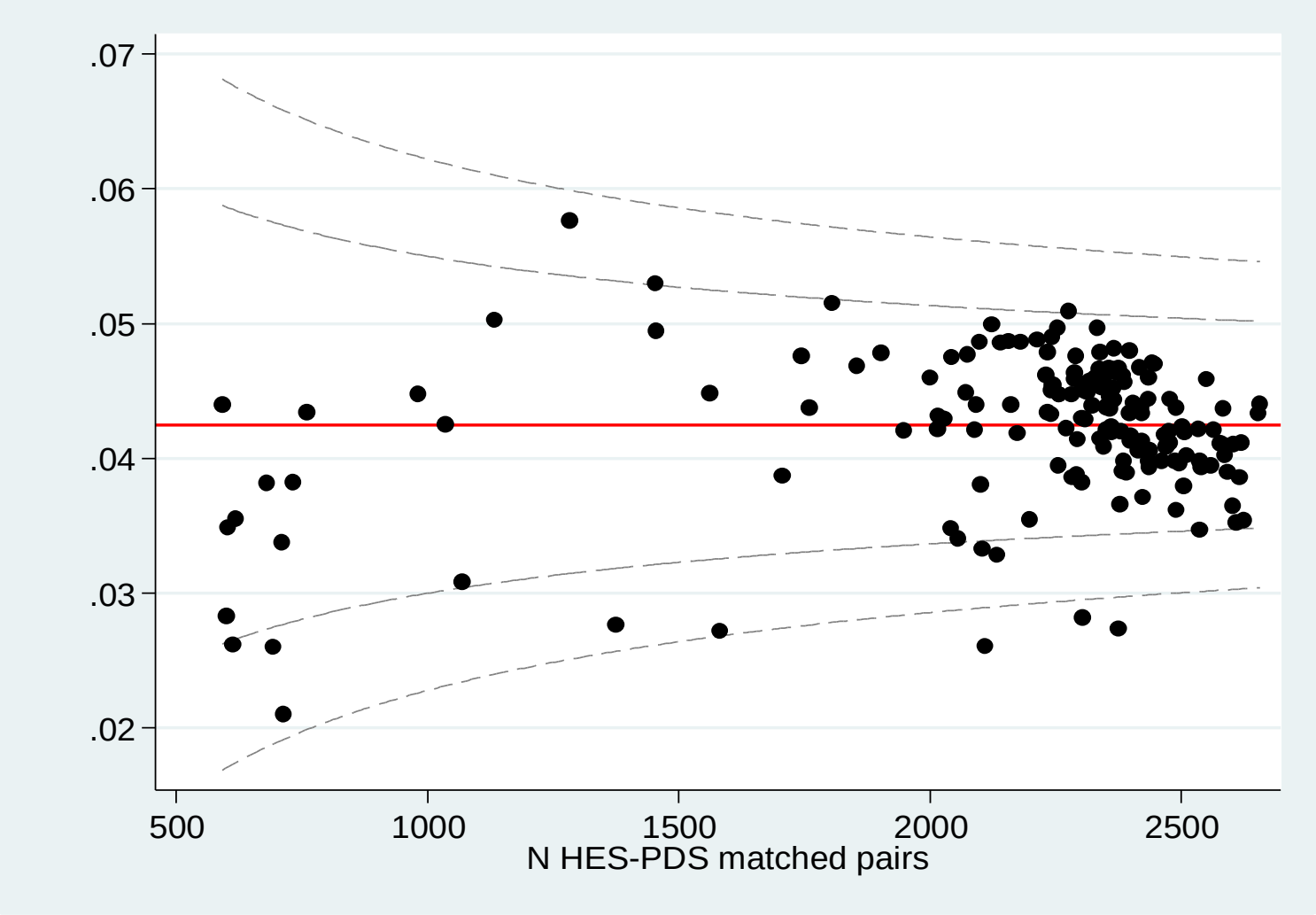
Ford JB, Roberts CL, Taylor LK (2006) Characteristics of unmatched maternal and baby records in linked birth records and hospital discharge data. Paediatr Perinat Ep 20 (4):329-337

# Variation - errors in identifiers in HES




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).

# Variation - errors in identifiers in HES




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).

# Differential (non-random) linkage – why?

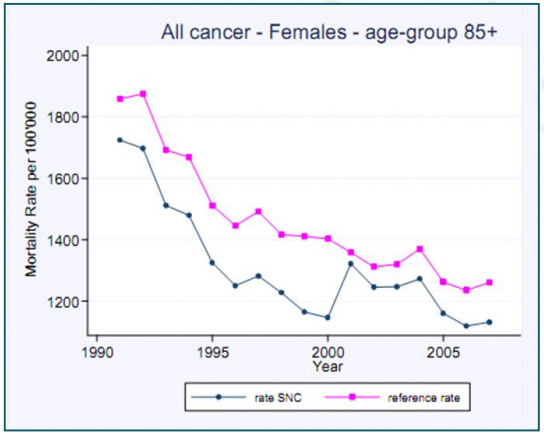
- Data quality differs by patient group / SES etc.
  - Bohensky et al 2010. Data Linkage: A powerful research tool with potential problems. *BMC Health Services Research*
- Unknown/estimated dates of birth
  - Unconscious, frail, dementia,
- Unconventional surnames
- Misleading information
  - Drug user,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

# Evaluating linkage

## i) 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 |

## ii) Gold-standard / reference data to quantify linkage errors




## iii) Sensitivity analysis using different probabilistic thresholds

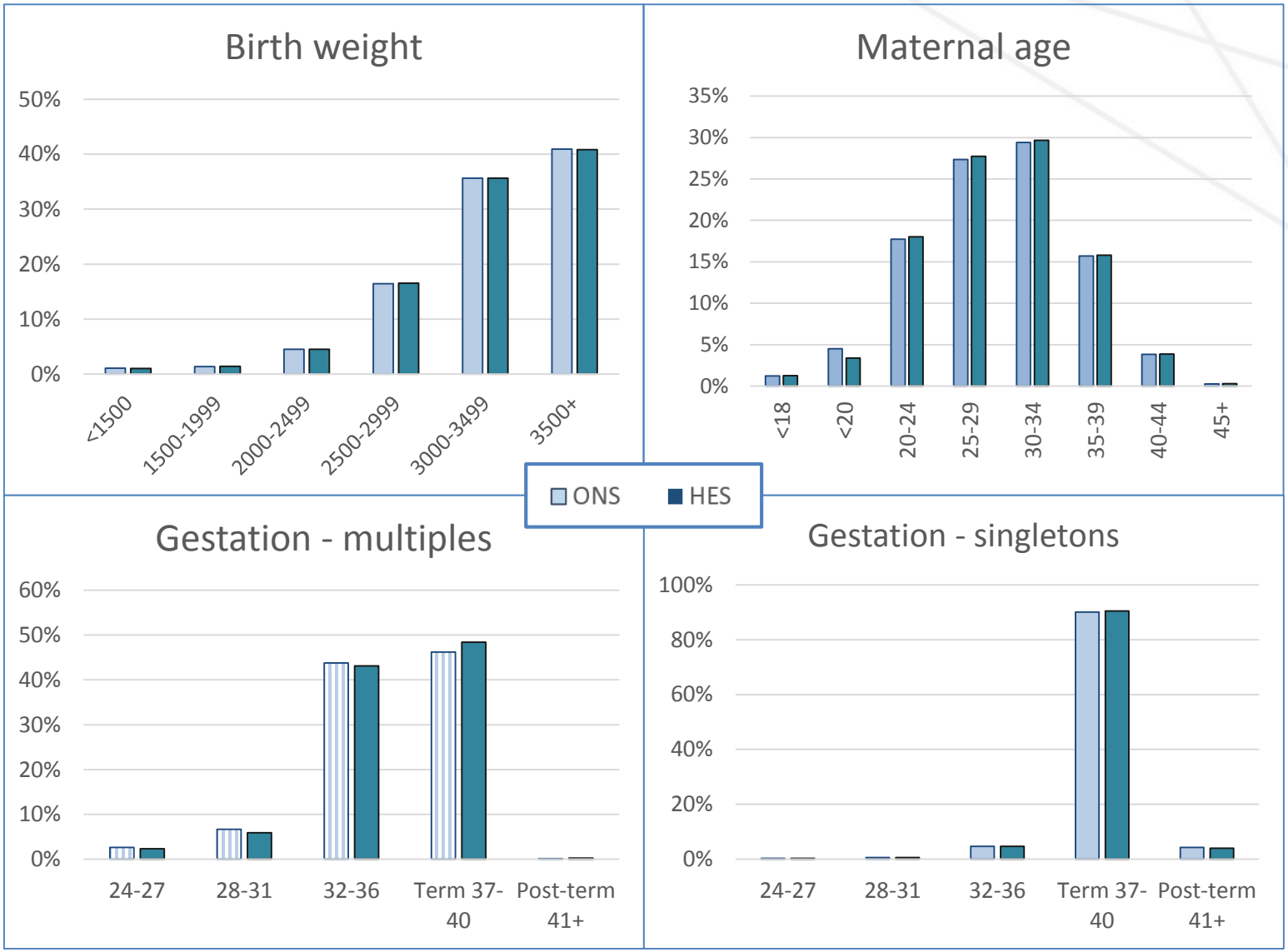
Highly sensitive

Highly specific

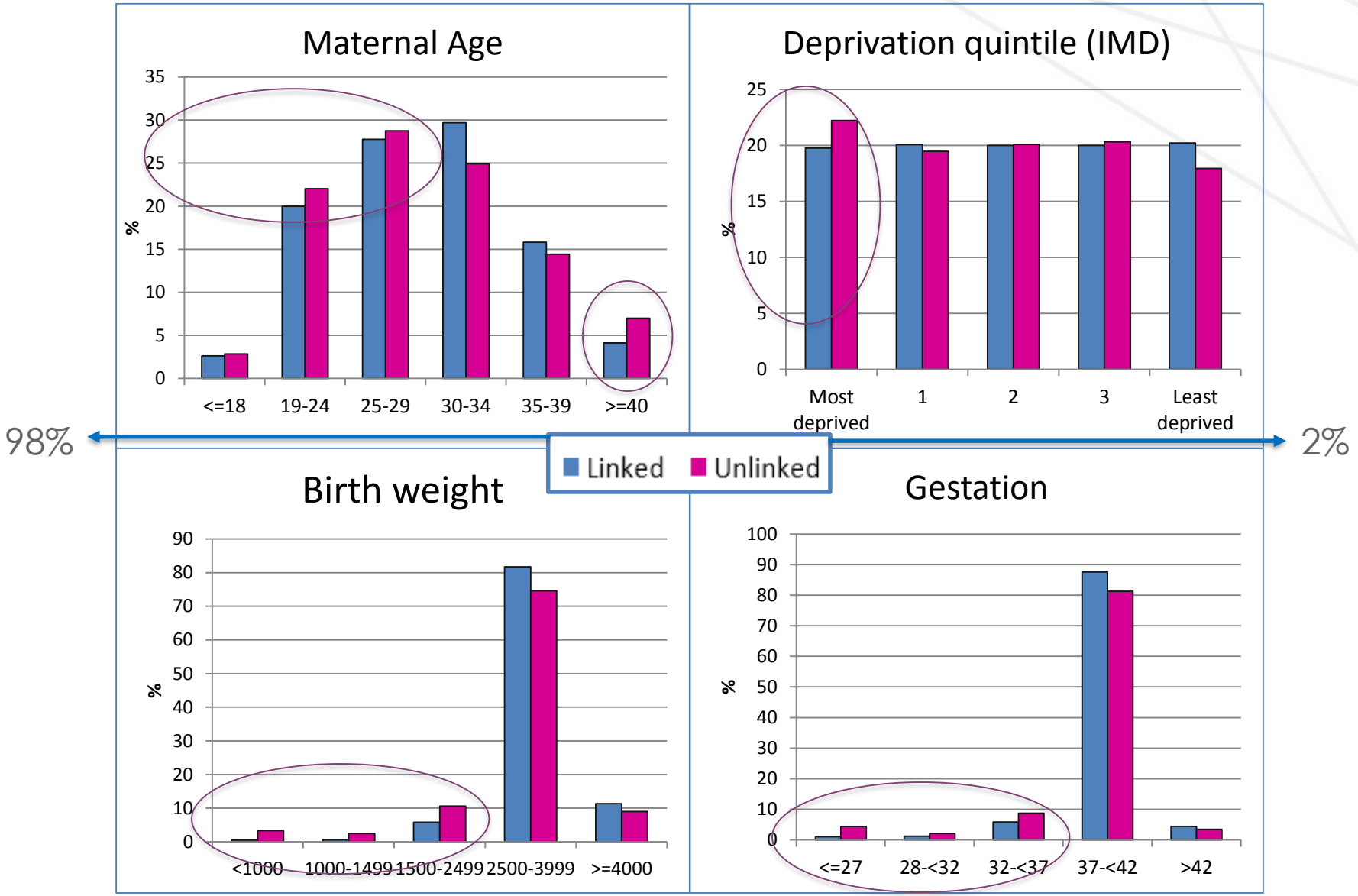
**Table 3.** Hazard Ratios for the Association Between Ethnicity and Mortality Using Three Linkage Criteria, 1989-2002

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

# Evaluating mother-baby linkage



# Evaluating mother-baby linkage



# Evaluating mother-baby linkage

## Gold-standard: 15 maternity units, 2012/13 (N=72,824)

### Original algorithm

- 632 (0.9%) false matches
- 297 (0.4%) missed matches

## Sensitivity analysis: Different linkage criteria

### Conservative prob algorithm

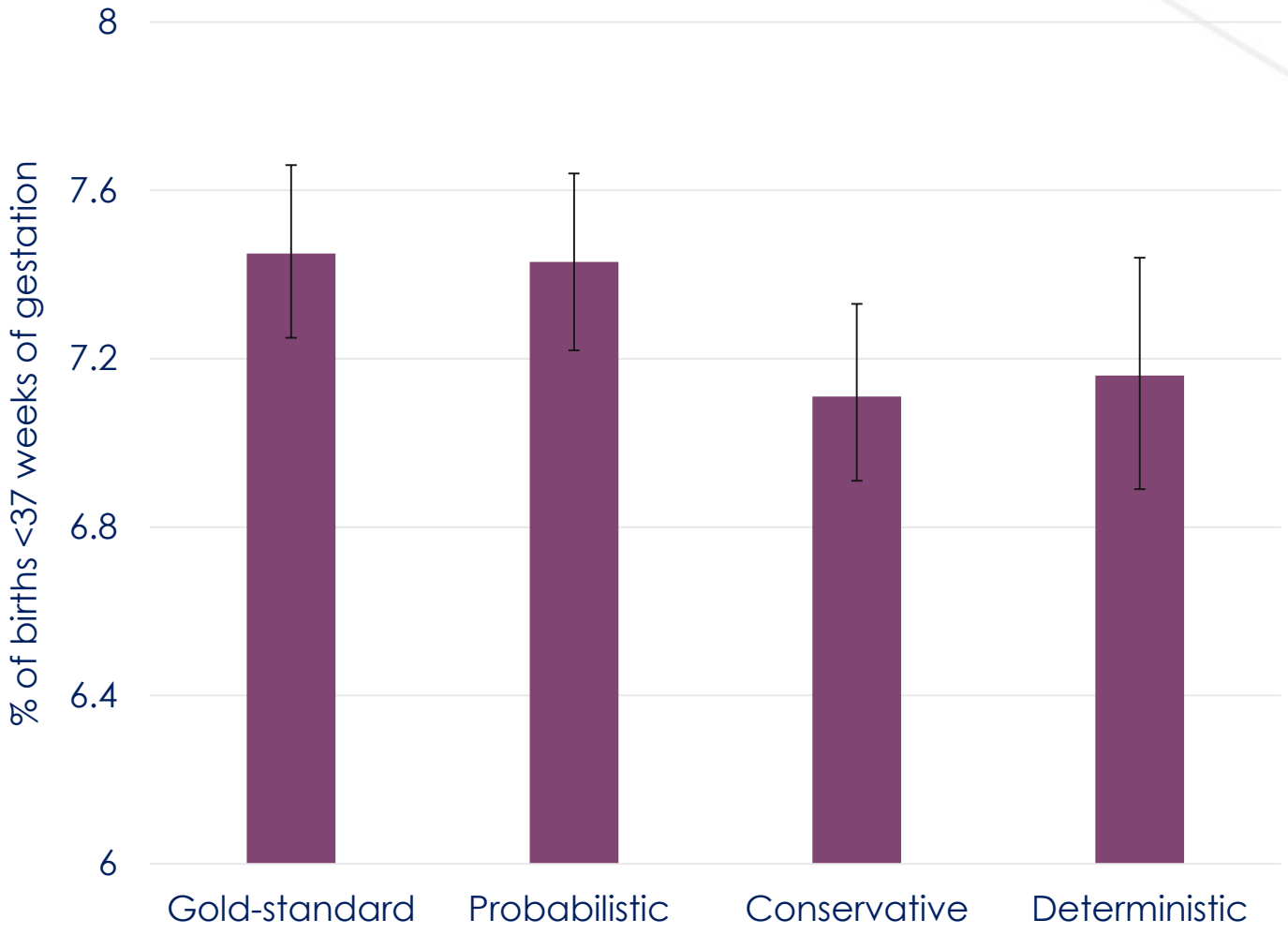
- 212 (0.3%) false matches
- 7,797 (10.7%) missed matches

### Deterministic only algorithm

- 22 (0.1%) false matches
- 37,515 (51.6%) missed matches

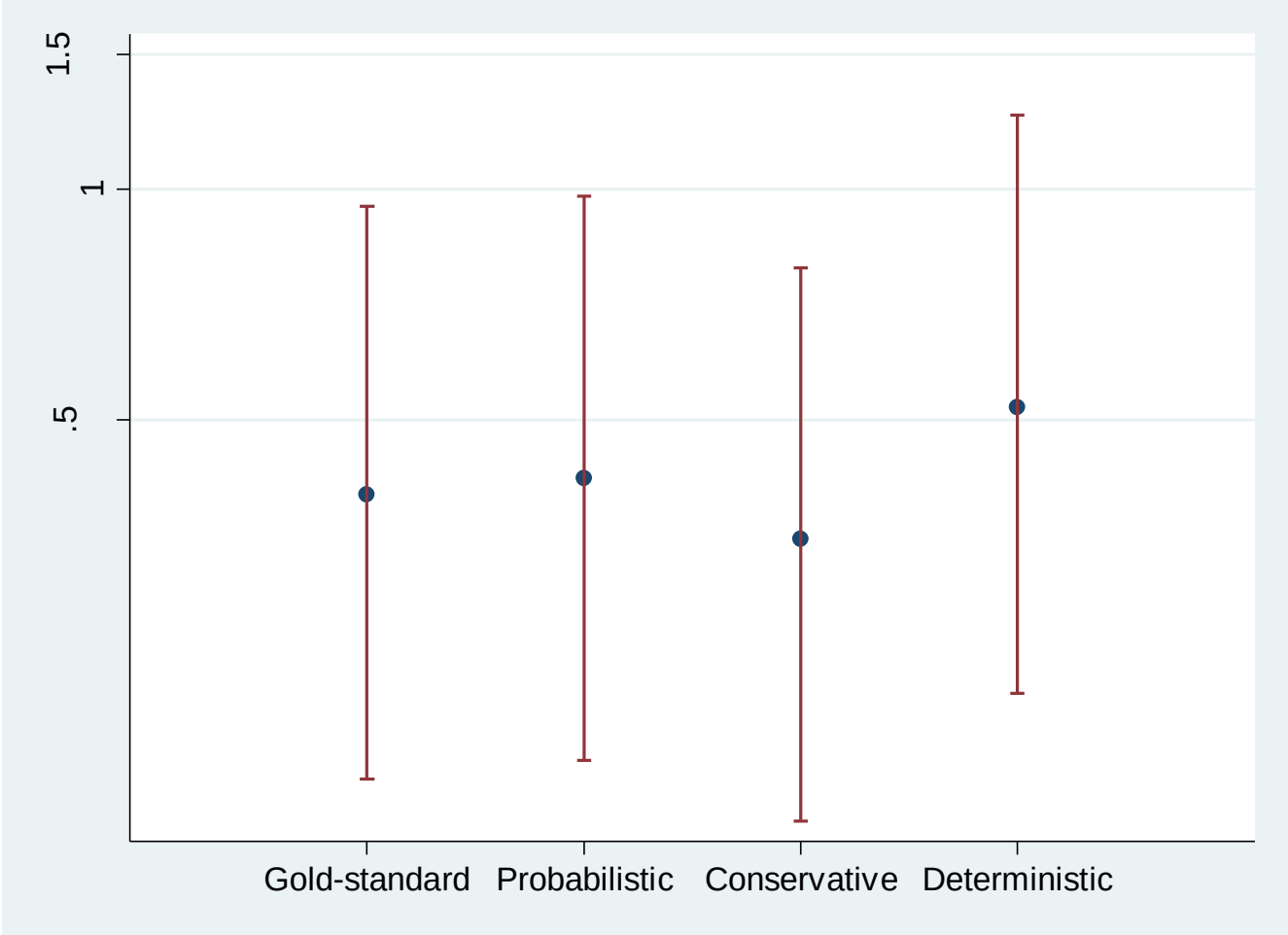
# Evaluating mother-baby linkage

Identifying impact on results: rate of preterm birth



# Evaluating mother-baby linkage

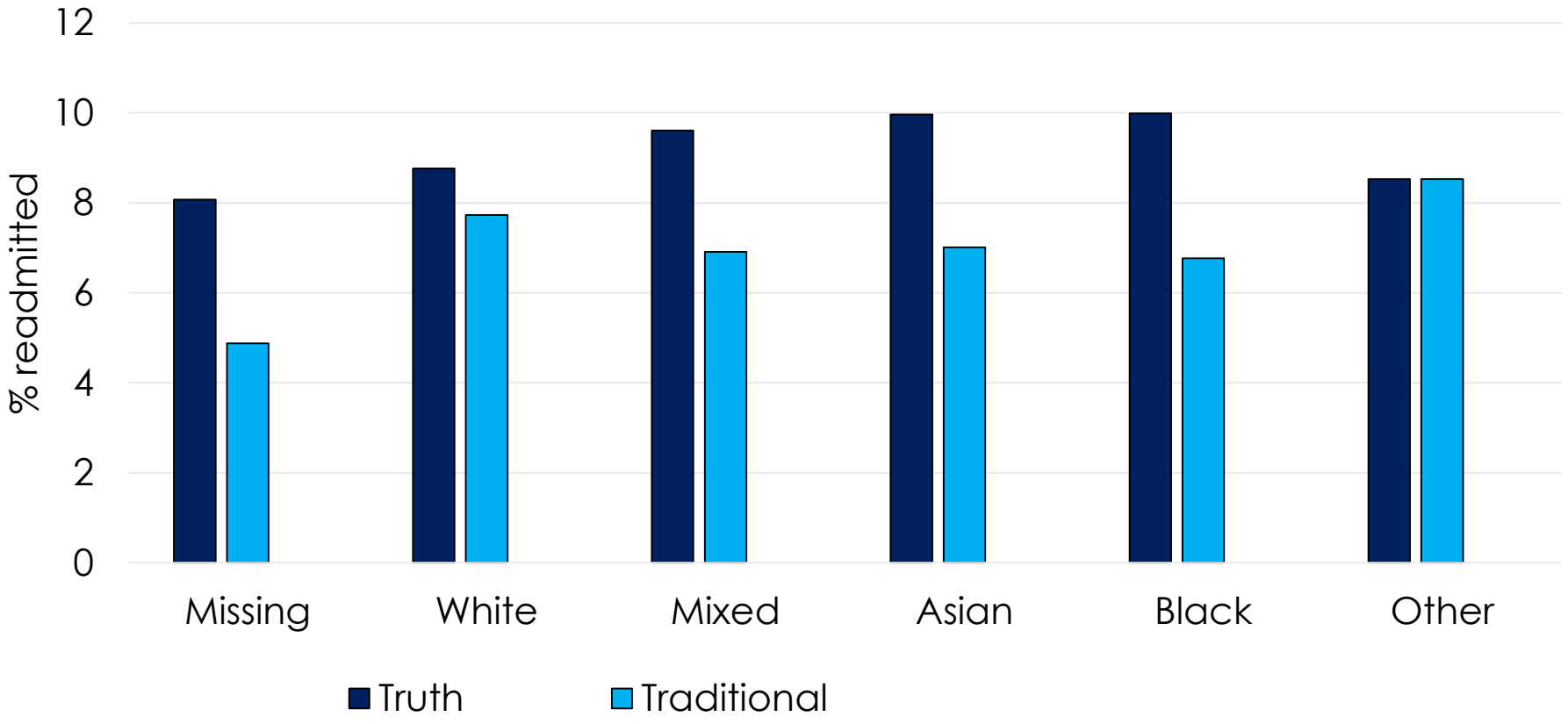
Identifying impact on results: association between maternal risk-factors and infant survival to discharge



# Alternative linkage methods

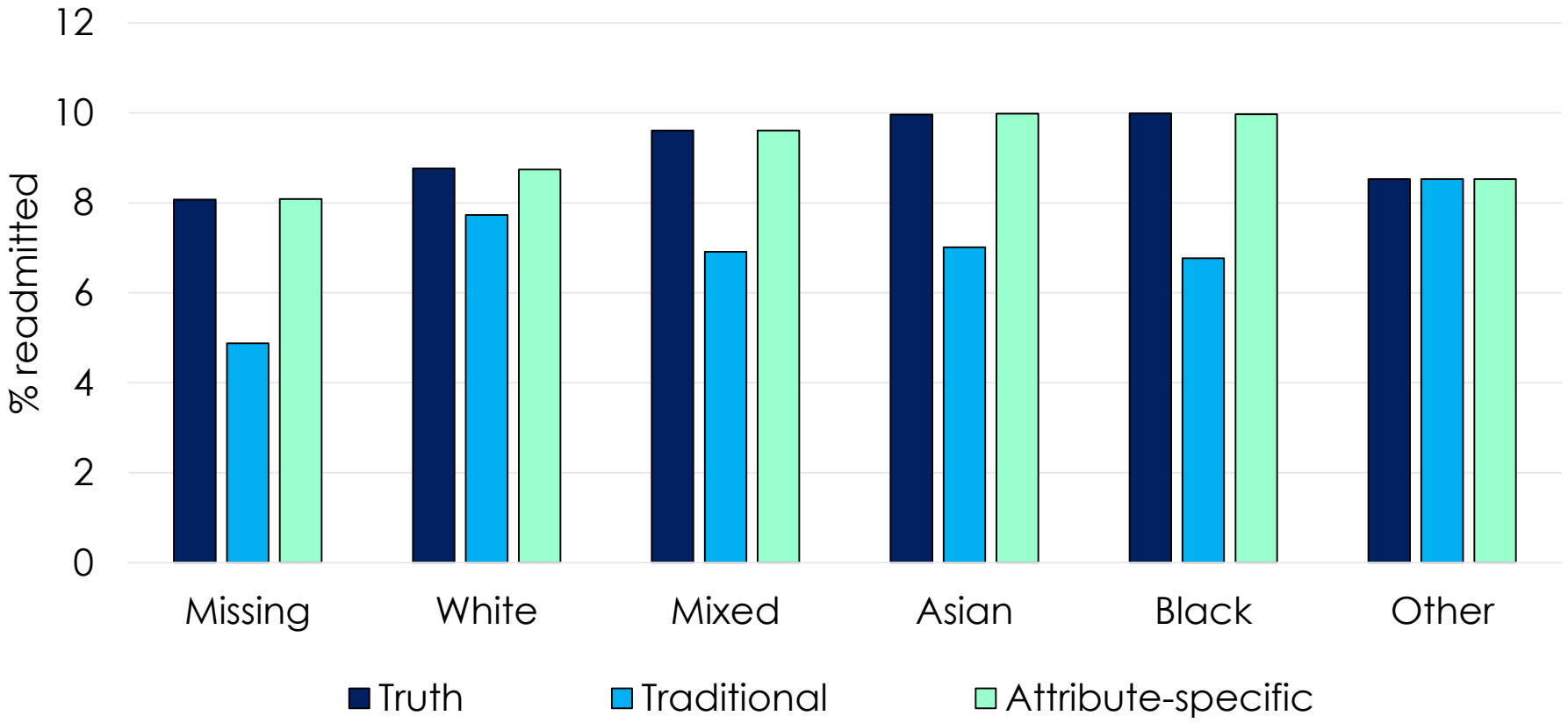
Attribute-specific match weights

- Exploiting what we know about variation in identifier errors
- E.g. when errors in identifiers differ between ethnic groups



# Alternative linkage methods

Attribute-specific match weights
- Exploiting what we know about variation in identifier errors
- E.g. when errors in identifiers differ between ethnic groups




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).

# Alternative linkage methods

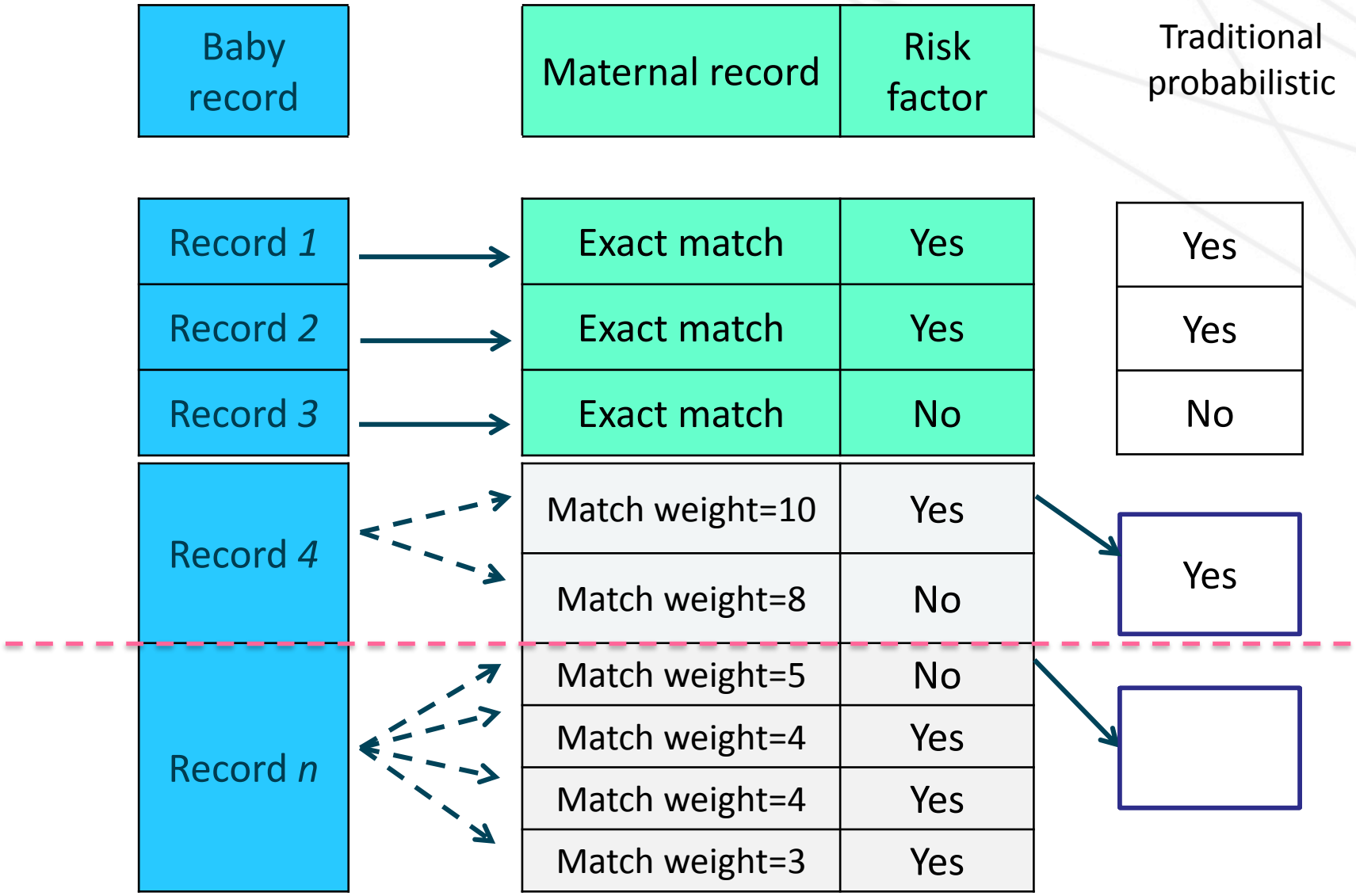
Calculation of match weights/scores without the need for training data

- Class of correspondence analysis
- Set constraints, e.g.
  - full agreement = 100,
  - no agreement = 0
- Given a set number of identifiers and levels of agreement between those identifiers, aim is to derive scores that minimise the total discrepancy within each pair of records

| | Day | Month | Year | Sex |
|---|---|---|---|---|
| **Scaling scores** | 53 | 22 | 19 | 7 |
| **Probabilistic weights** | 32 | 27 | 26 | 15 |

Goldstein, H., Harron, K. & Cortina-Borja, M. A scaling approach to record linkage. *Stat Med* **accepted** (2017).

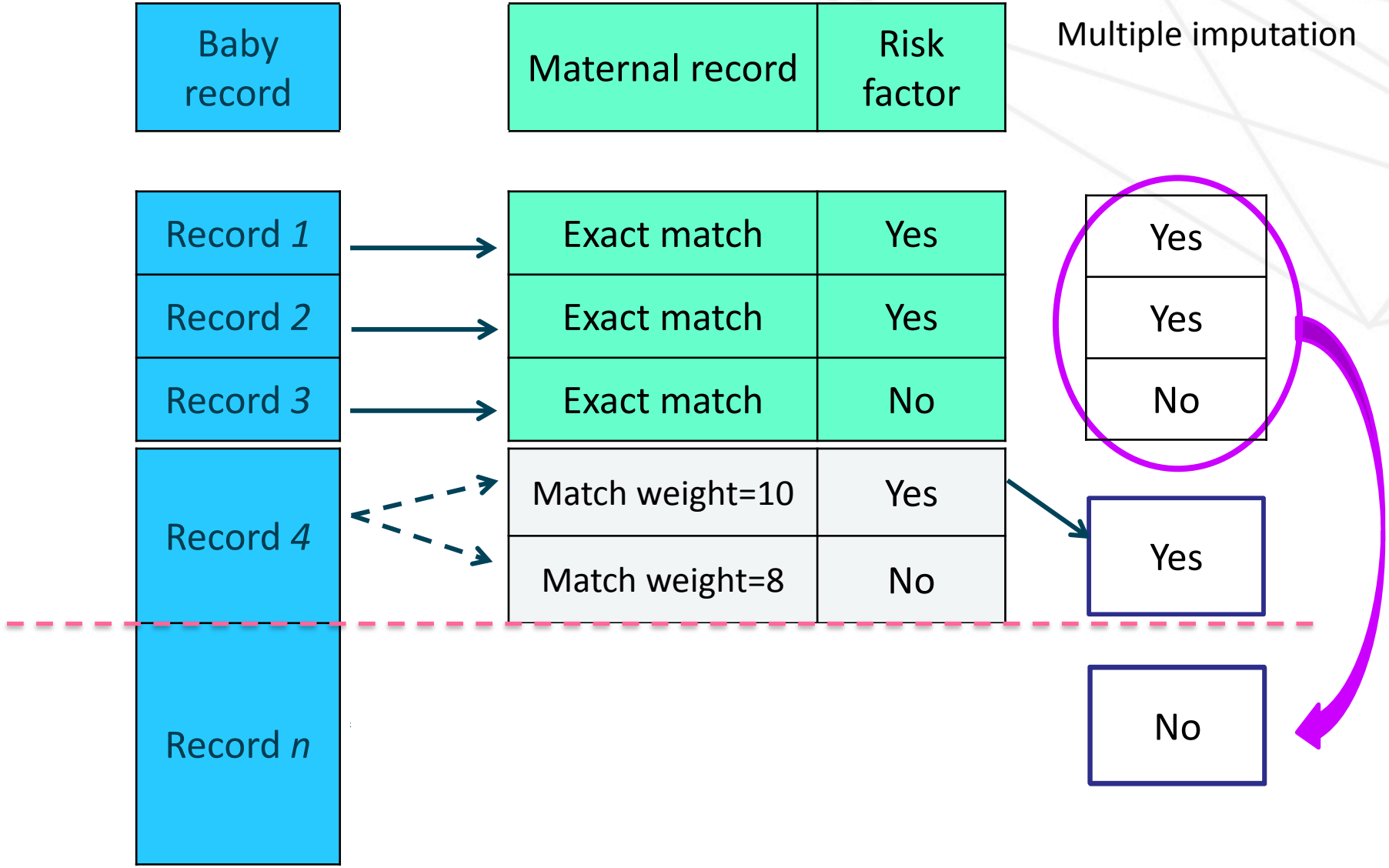
# Accounting for linkage error




Goldstein et al *Stat Med* 2012;31(28):3481-3493

Harron et al *BMC Med Res Method* 2014;14(1):36

# Imputation for missed links

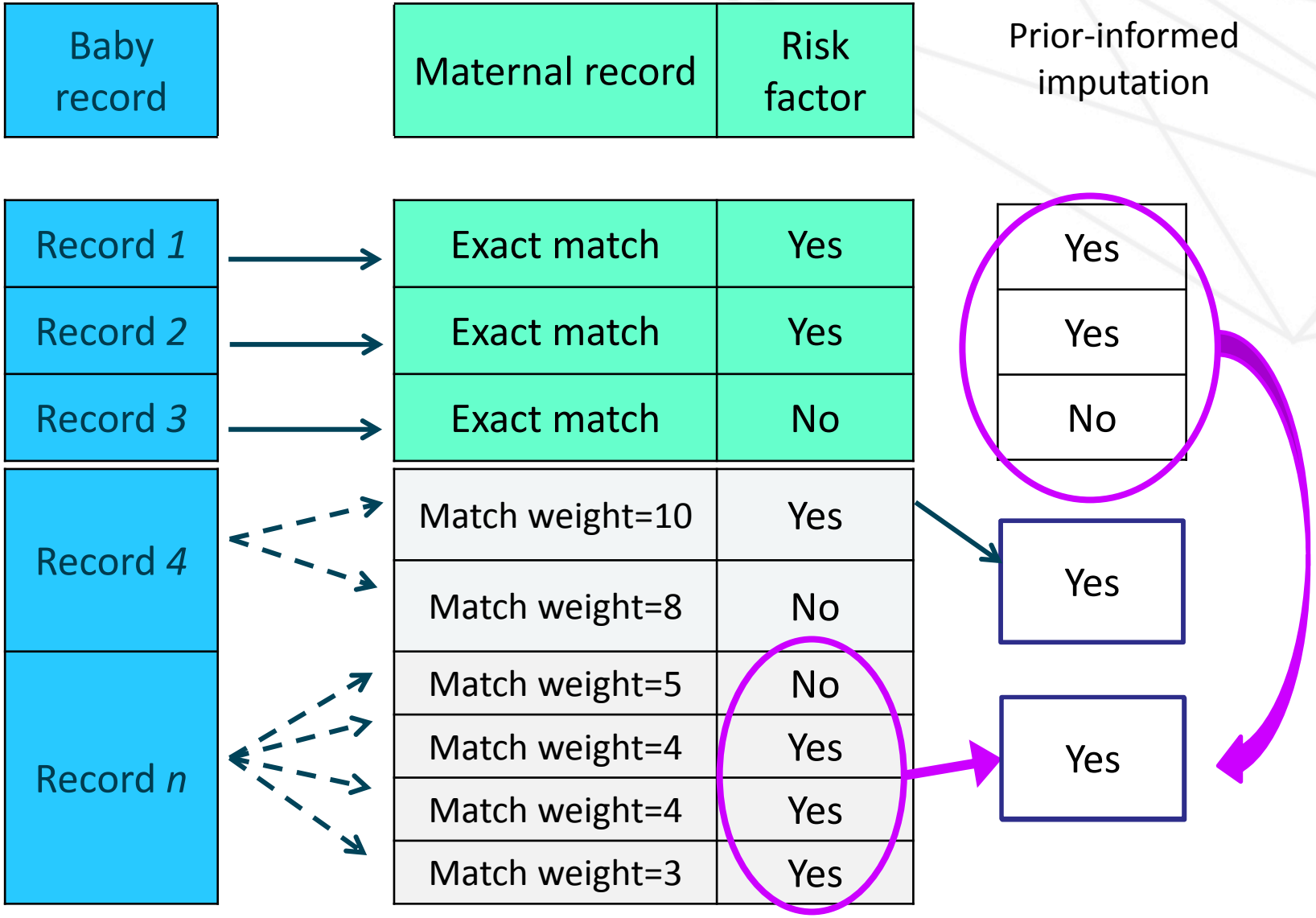




Goldstein et al *Stat Med* 2012;31(28):3481-3493

Harron et al *BMC Med Res Method* 2014;14(1):36

# Imputation for linkage uncertainty





Goldstein et al *Stat Med* 2012;31(28):3481-3493

Harron et al *BMC Med Res Method* 2014;14(1):36

# 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 (difficult with trusted third party model)

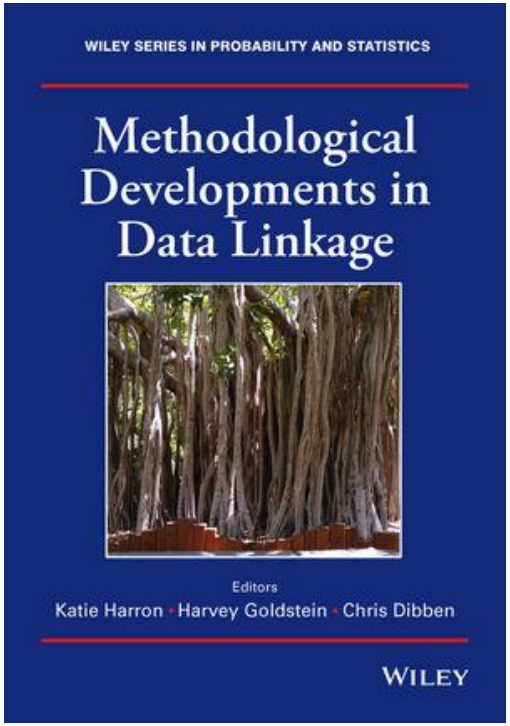
- Unfulfilled opportunities
  - Linkage between health and other sectors
  - Linkage of trial data for long-term follow up / safety analyses

# 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)



Administrative Data Research Centre England

Farr The Farr Institute London

wellcome trust Fellow

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE