



**GREAT ORMOND STREET
INSTITUTE OF CHILD HEALTH**

icnarc | intensive care
national audit &
research centre



QUANTITATIVE BIAS ANALYSIS TO ACCOUNT FOR LINKAGE ERROR

James Doidge¹, Katie Harron²

¹ Intensive Care National Audit and Research Centre (ICNARC);

^{1,2} UCL Great Ormond Street Institute of Child Health

James.Doidge@icnarc.org

Acknowledgements

Co-authors and collaborators:

- Joan Morris, Katie Harron, Sarah Stevens, Ruth Gilbert

Data collectors and providers, in particular:

- Public Health England National Congenital Anomaly and Rare Disease Registration Service (NCARDRS) and the data notifiers without whom no data would be available.

Funders:

- Economic and Social Research Council
- NIHR Great Ormond Street Hospital Biomedical Research Centre
- Farr Institute of Health Informatics Research
- Health Data Research UK

Aims

- Construct a **population birth cohort** of children with Down's syndrome and matched controls, for research on epidemiology and health and education services
- Establish **methods** for using linked administrative & registry data that can be extended to other data sources, and other rare diseases or congenital anomalies

Objectives

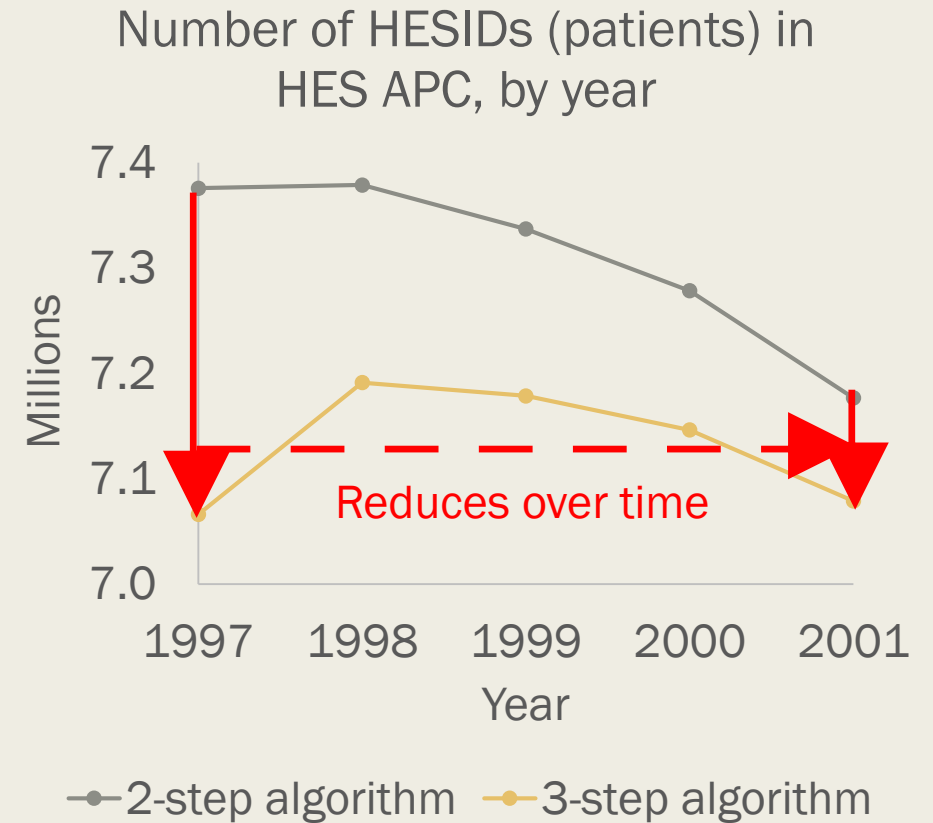
- Estimate the **prevalence** of Down's syndrome among live births in England, using two linked datasets:
 - *National Down Syndrome Cytogenetic Register (NDSCR; now part of NCARDRS)*
 - *Hospital Episode Statistics for England (HES)*
- Assess **population coverage/detection** rates in each data source

Options for estimating prevalence

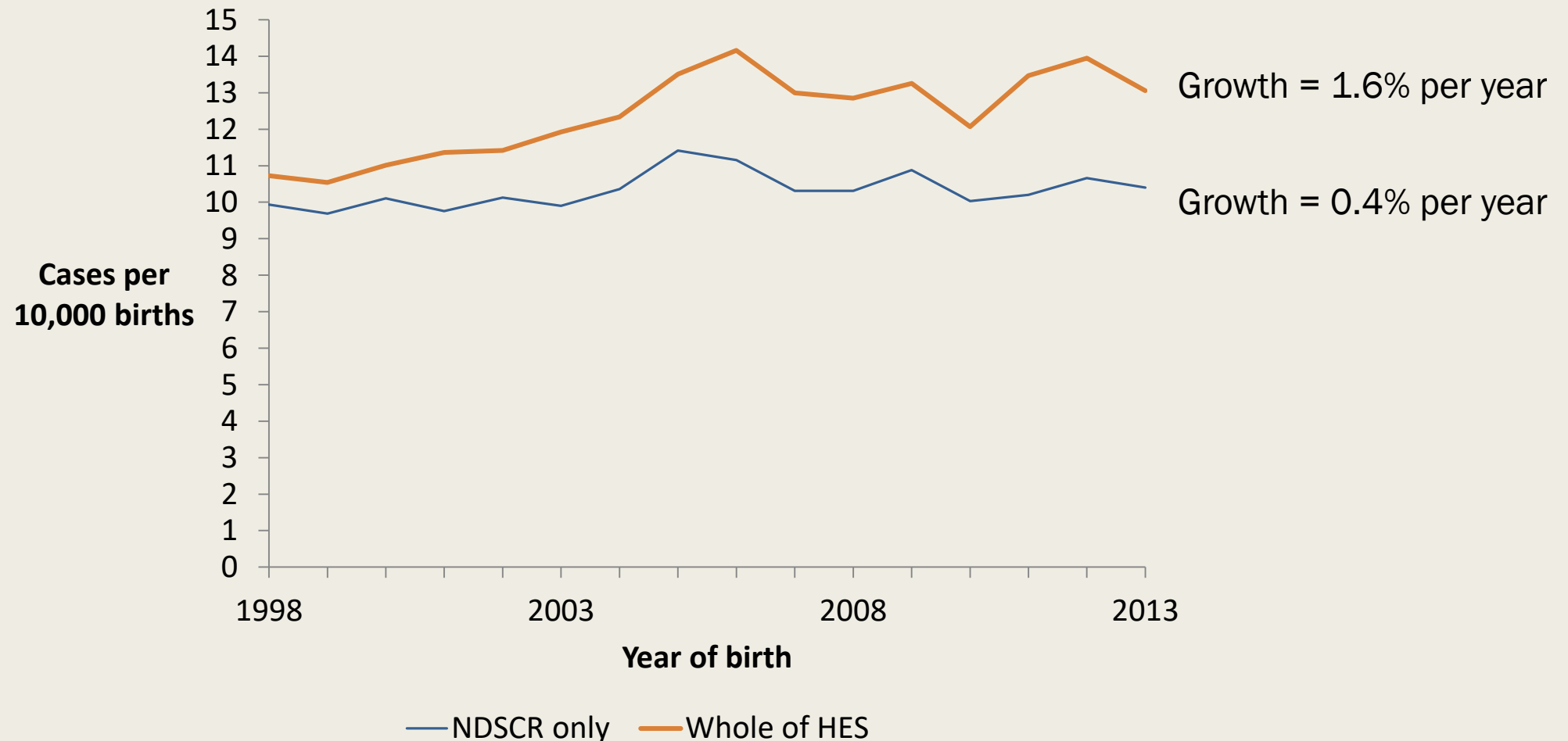
Using register data only (NDSCR)	Numerator only; requires denominator from ONS register of births
Using hospital data only (HES)	(i) All HESIDs with DOB within target period (ii) Administrative birth cohort: HESIDs with identified birth episodes
Using linked data	(i) Using HES to define population, with supplementary diagnostic information from linked NDSCR records (ii) Using pooled data +/- capture-recapture analysis

Known issues with HES internal linkage (HESID)

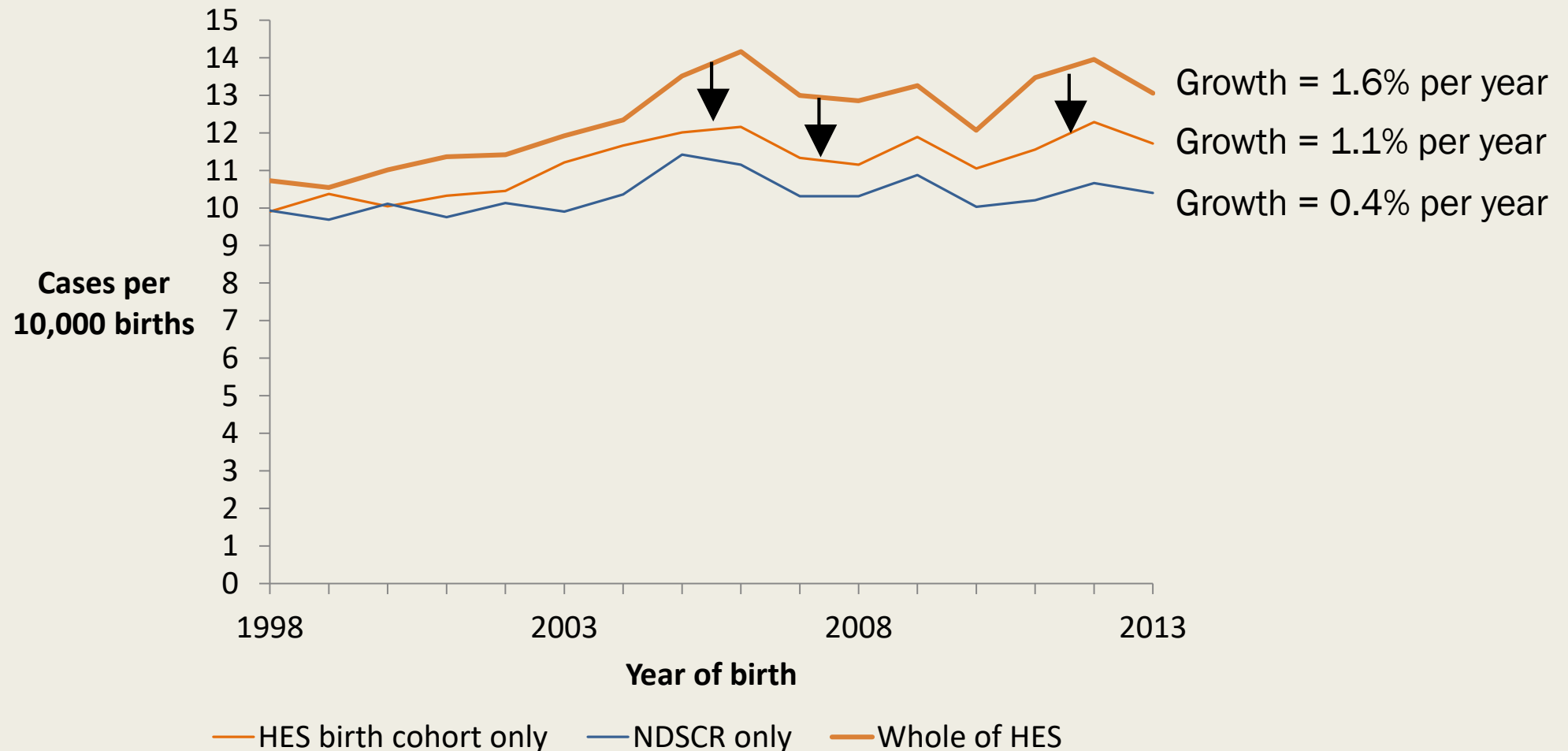
- Allocation of HESIDs is deterministic
 - *Relies heavily on NHS numbers, which are substantially incomplete prior to 2009*
- Prior to 2002, NHS numbers were allocated at GP registration (**not birth**)
- Many birth episodes allocated unique HESIDs
 - *Missing links to subsequent episodes*
 - *'Splitting' of people (esp. babies) into multiple HESIDs → double-counting*
- False links and 'merging' (multiple people sharing a HESID) also known to occur



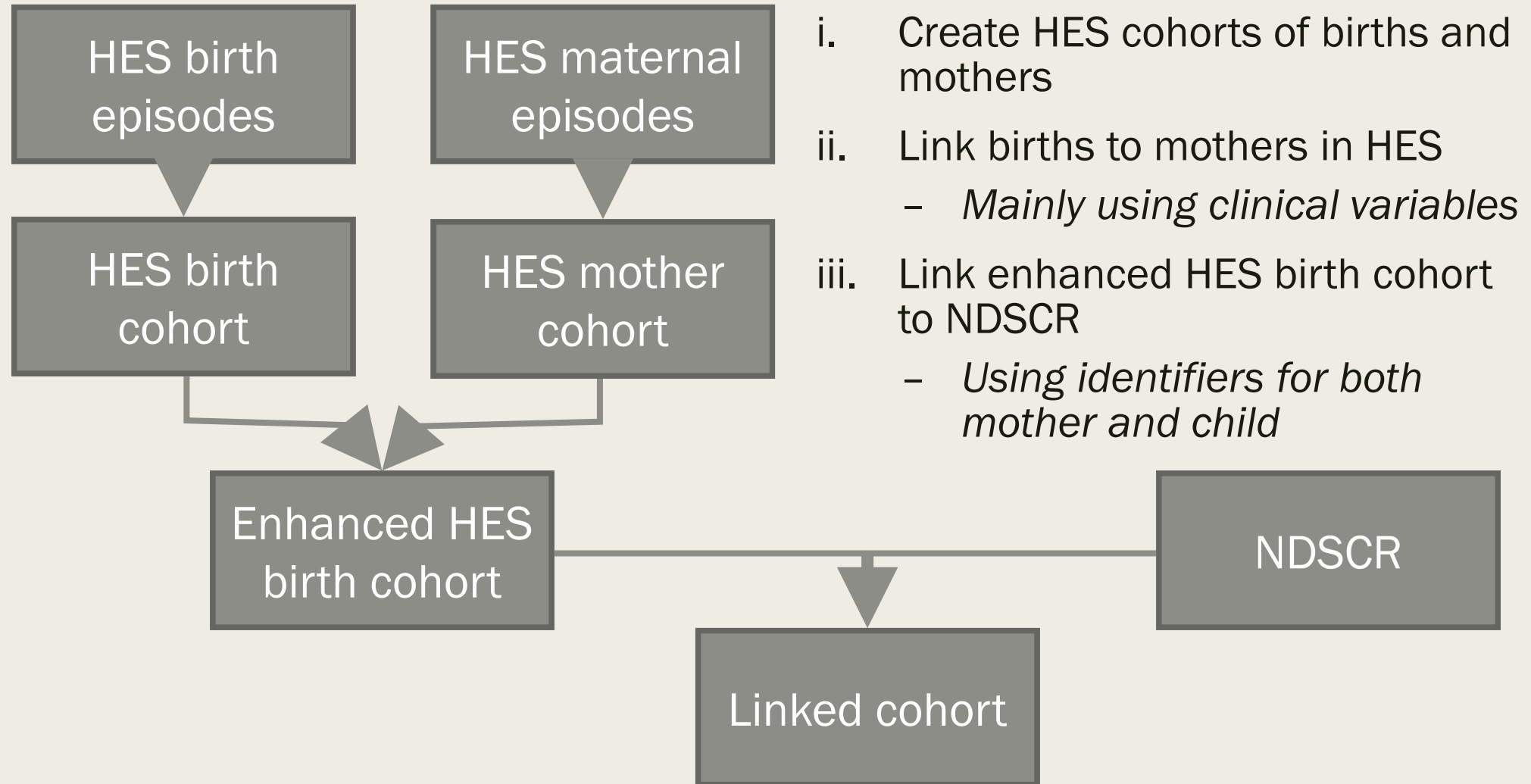
Estimated prevalence: NDSR-only vs HES-only (all HESIDs)



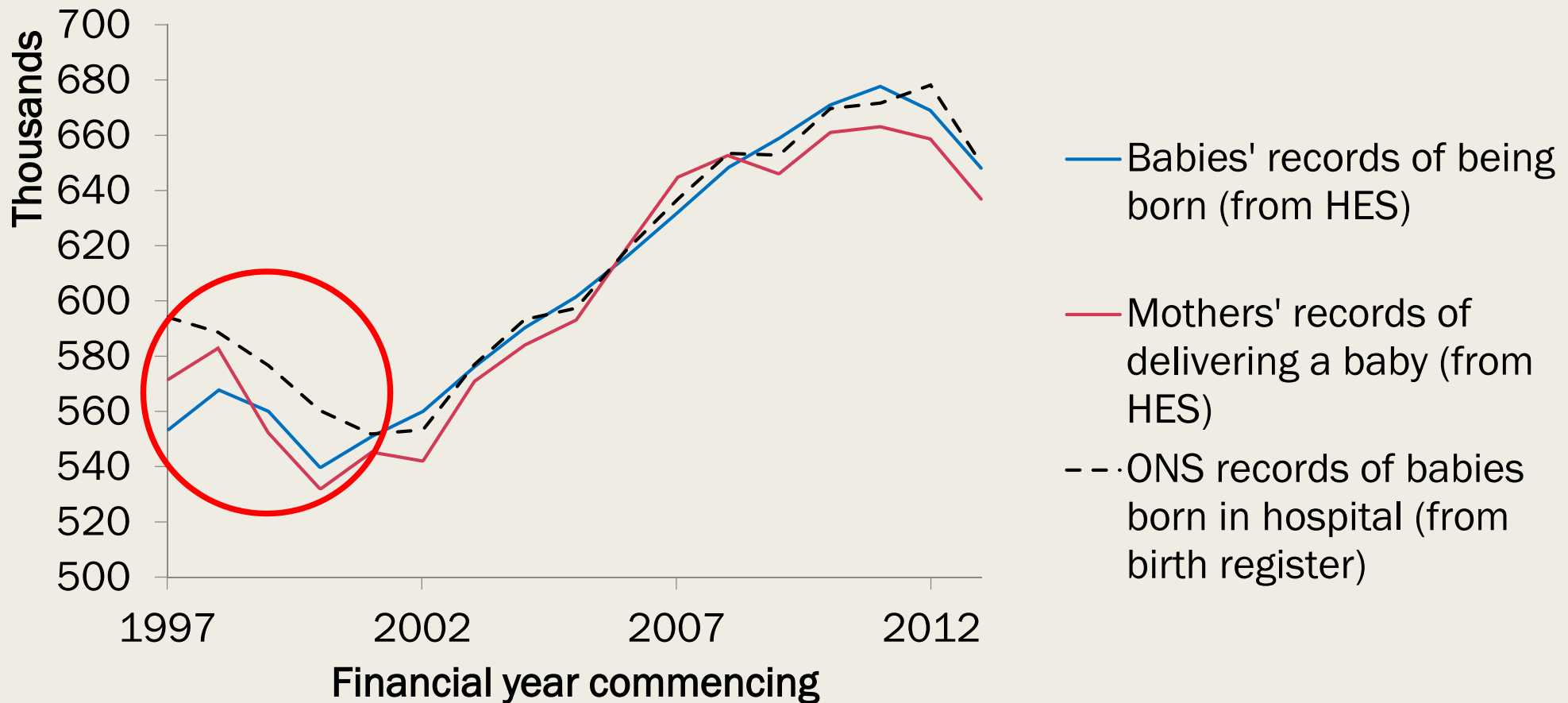
Estimated prevalence: Restricting HES to a birth cohort



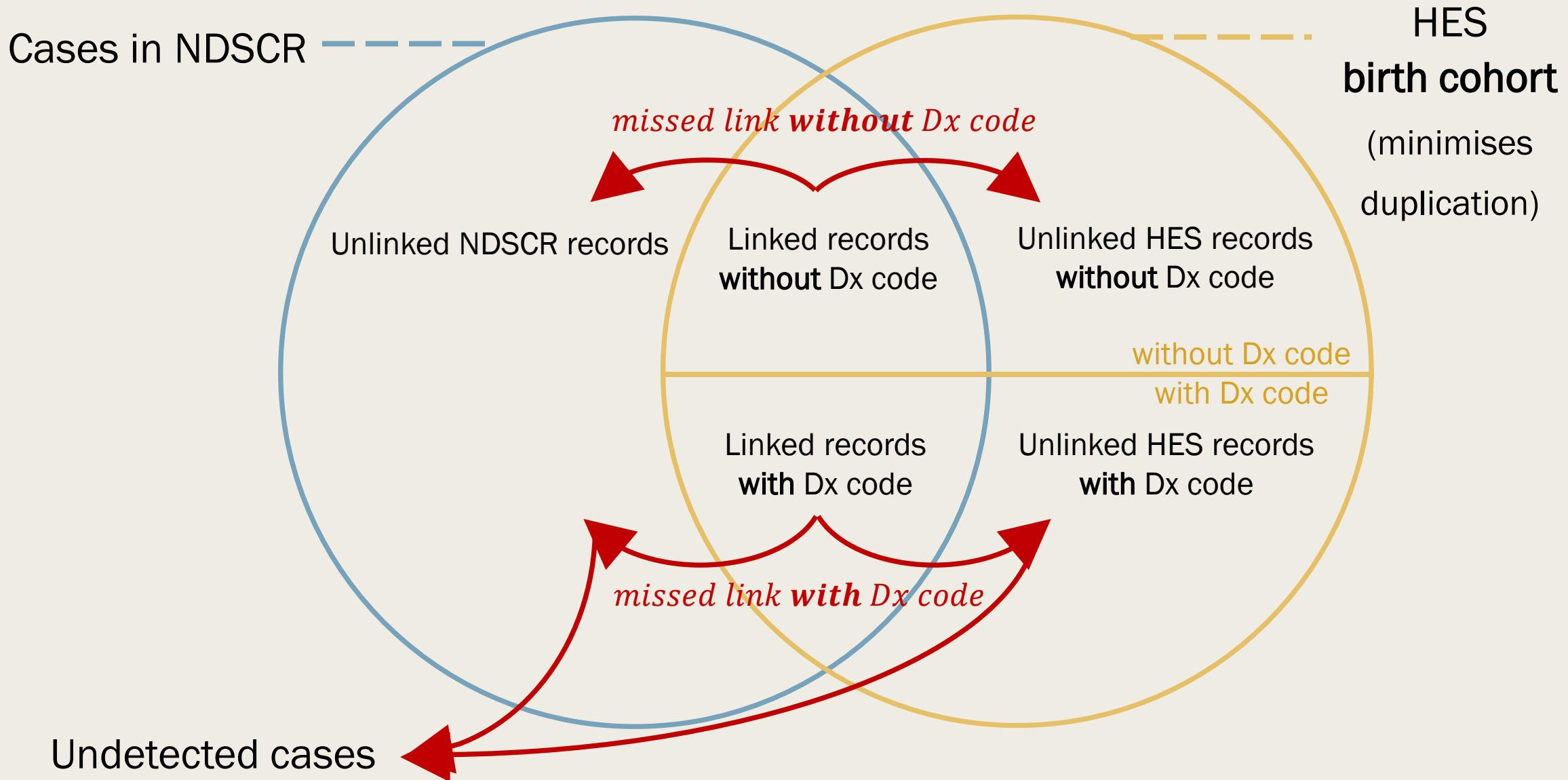
NDSCR-HES linkage overview



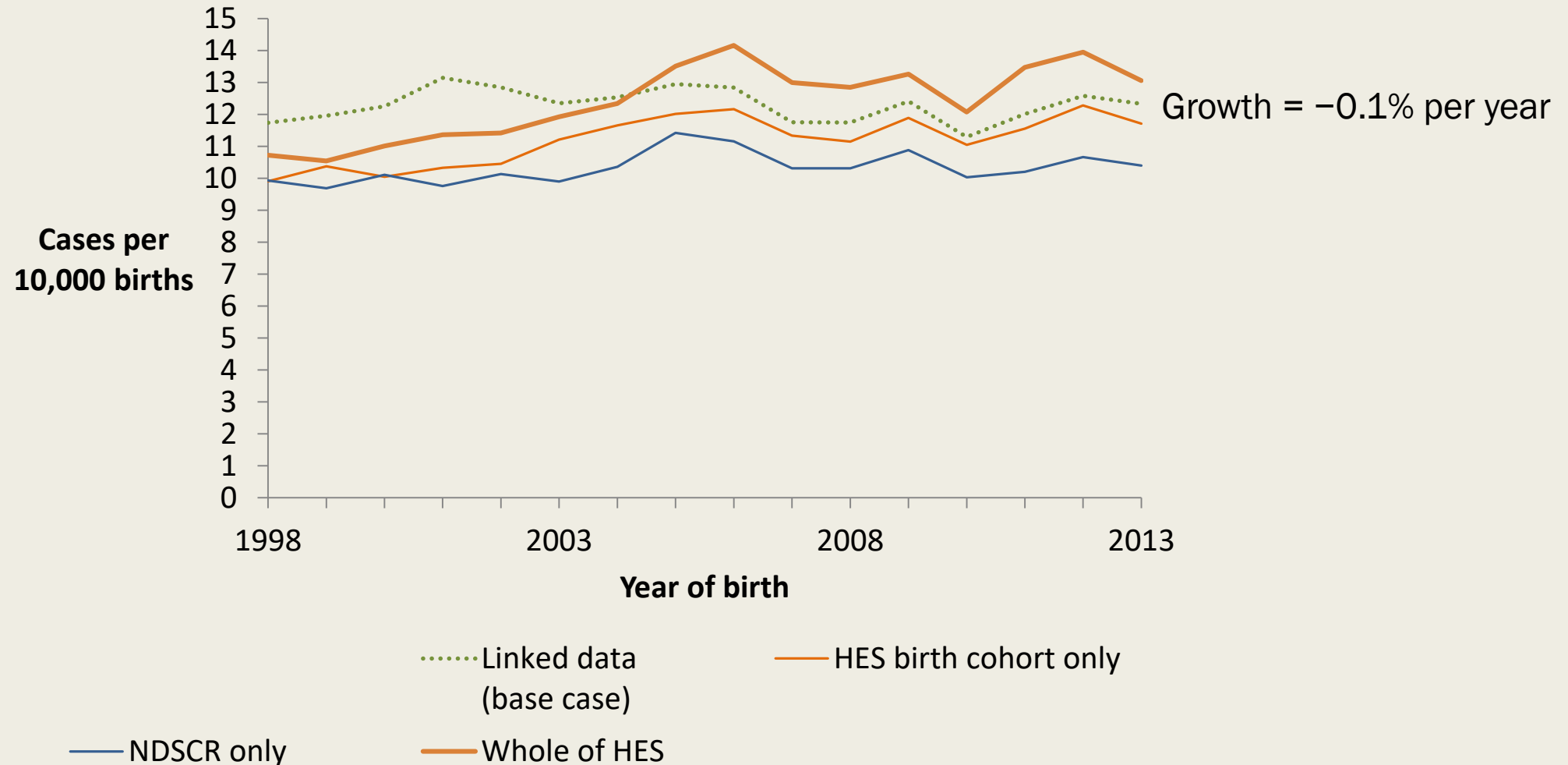
Comparison to external reference statistics (HES mother & baby cohorts)



Potential manifestations of linkage error



Estimated prevalence: Using linked data



Comparison of linked/unlinked records (and high/low agreement)

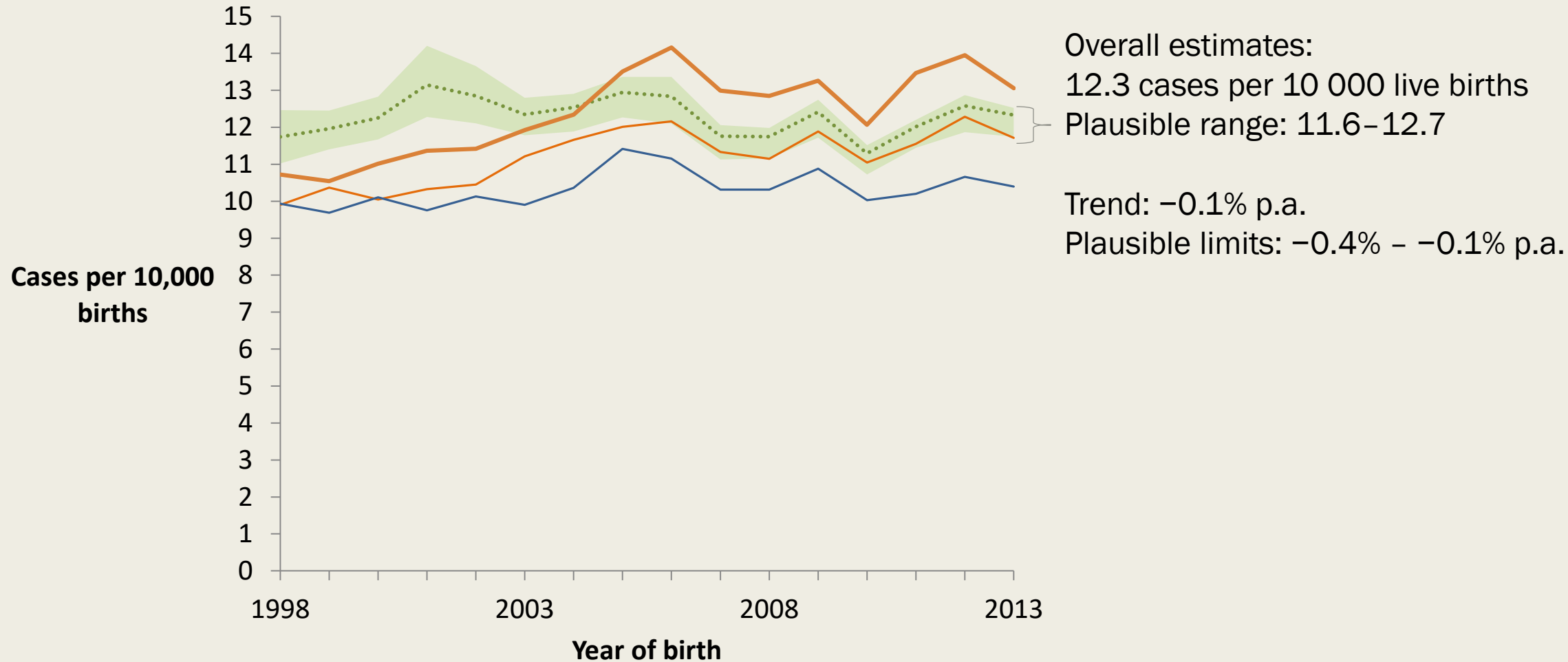
	Deterministic links	Probabilistic (MW > 40.6)	Probabilistic (MW: 30.5–40.6)	Probabilistic (MW: 18.1–30.5)	Probabilistic (MW < 18.1)	Unlinked NDSCR	Unlinked HES
Q90 Dx code	96.4%	91.1%	70.2%	81.4%	17.0%	—	—
Difference in DOB > 180 days	0.4%	< 0.3%	1.5%	3.6%	6.0%	—	—
Age at diagnosis ≥ 1yr							
<i>in NDSCR records</i>	0.6%	0.3%	1.0%	2.6%	10.1%	7.4%	—
<i>in HES records</i>	9.1%	10.2%	9.6%	11.8%	11.1%	—	22.3%
Number of episodes in first year of life (in HES records)							
1	22.5%	38.4%	48.6%	42.4%	78.2%	—	36.1%
2–4	42.5%	37.1%	30.4%	31.5%	15.4%	—	34.5%
≥ 5	35.0%	24.4%	20.9%	26.1%	6.5%	—	29.4%

Quantitative bias analysis

- Assigned base case estimates + plausible limits for key bias parameters:

Bias parameter	Lower limit	Base case	Upper limit
Proportion of links that are true (precision) varied by match weight:			
<i>Deterministic + clerical review</i>	100%	100%	100%
<i>Probabilistic (match weight > 40.6)</i>	99%	100%	100%
<i>Probabilistic (match weight: 30.5–40.6)</i>	95%	98%	100%
<i>Probabilistic (match weight: 18.1–30.5)</i>	80%	90%	100%
<i>Probabilistic (match weight: 0.0–18.1)</i>	50%	80%	100%
Proportion of unlinked NDSCR records that are missed links	10%	50%	90%
Positive predictive value of diagnosis codes among unlinked HES cases	95%	99.5%	100%

Using linked data with quantitative bias analysis



Conclusions

1. Coverage of *live births* in NDSCR improved slowly over time
 - *Prenatal diagnoses with unknown birth outcomes difficult to link*
2. Trends in HES primarily reflects improving quality of HESIDs (even with birth cohort approach)
3. Live birth prevalence of Down's syndrome appears stable

Take-home points

1. Trends in administrative data can be highly susceptible to variation in data quality over time
2. Linkage error within datasets can really screw up analysis
3. Linkage can be used to assess data quality of both datasets and provide more robust analysis than either, even when there is uncertainty in linkage.
4. Linkage error bias analysis is always possible, even without evidence of error rates, once you work out how linkage error will manifest in an analysis.

Outputs

- Main paper in press (IJPDS)
- Conceptual paper on linkage error bias analysis published in IJE



International Journal of Epidemiology, 2019, 1–11
doi: 10.1093/ije/dyz203
Education Corner



Education Corner

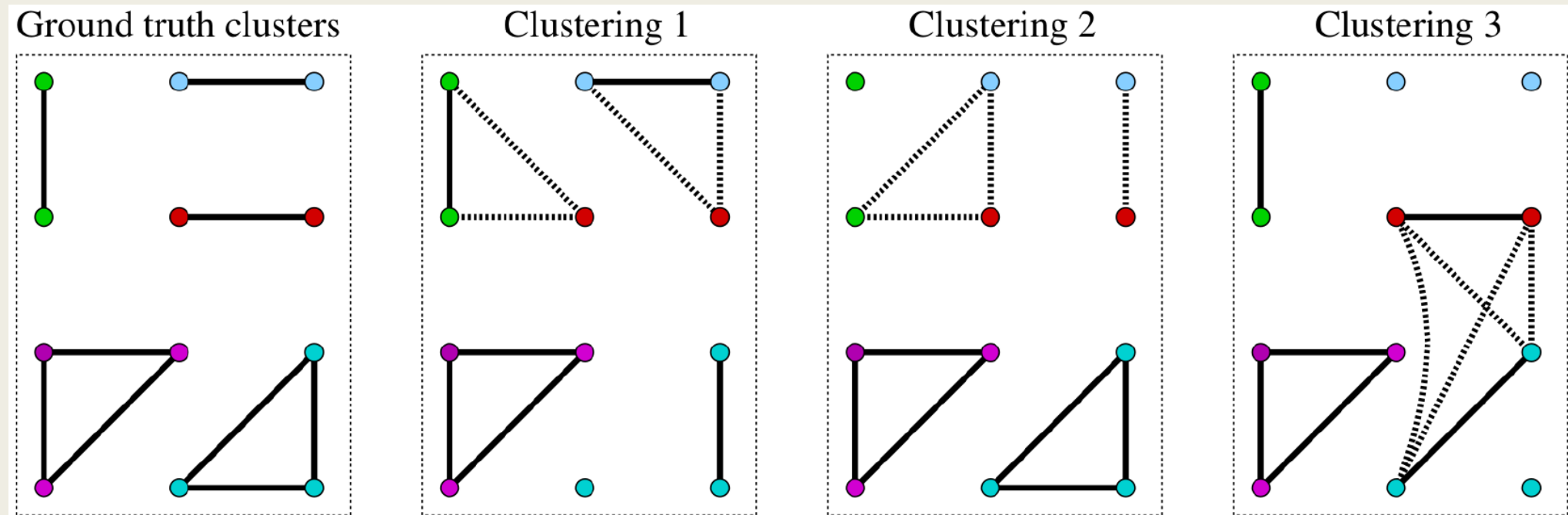
**Reflections on modern methods:
linkage error bias**

James C Doidge ^{1,2*} and Katie L Harron ²

- Look out for the forthcoming National Statistician's Quality Review on Data Linkage!

Further problems and future directions

- Probabilistic bias analysis for linkage error
- Imputation methods for many:many linkage/unknown clusters



Nanayakkara C, Christen P, Ranbaduge T, Garrett E. Evaluation measure for group-based record linkage. International Journal of Population Data Science. 2019;4. doi: 10.23889/ijpds.v4i1.1127.