Prevalence of Down's Syndrome in England, 1998–2013:
Comparison of linked surveillance data and electronic health records

# Abstract

## Introduction

Disease registers and electronic health records are valuable resources for disease surveillance and research but can be limited by variation in data quality over time. Quality may be limited in terms of the accuracy of clinical information, of the 'internal linkage' that supports person-based analysis of most administrative datasets, or by errors in linkage between multiple datasets.

## Objectives

By linking the National Down Syndrome Cytogenetic Register (NDSCR) to Hospital Episode Statistics for England (HES), we aimed to assess the quality of each and establish a consistent approach for analysis of trends in prevalence of Down’s syndrome among live births in England.

## Methods

Probabilistic record linkage of NDSCR to HES for the period 1998–2013, supported by linkage of babies to mothers within HES. Comparison of prevalence estimates in England using NDSCR only, HES data only, and linked data. Capture-recapture analysis and quantitative bias analysis were used to account for potential errors, including false positive diagnostic codes, unrecorded diagnoses, and linkage error.

## Results

Analyses of single-source data indicated increasing live birth prevalence of Down’s Syndrome, particularly analysis of HES. Linked data indicated a contrastingly stable prevalence of 12.3 (plausible range: 11.6–12.7) cases per 10 000 live births.

## Conclusions

Case ascertainment in NDSCR improved slightly over time, creating a picture of slowly increasing prevalence. The emerging epidemic suggested by HES primarily reflects improving linkage within HES (assignment of unique patient identifiers to hospital episodes). Administrative data are valuable but trends should be interpreted with caution, and with assessment of data quality over time. Data linkage with quantitative bias analysis can provide more robust estimation and, in this case, stronger evidence that prevalence is not increasing. Routine linkage of administrative and register data can enhance the value of each.

Keywords
Down’s syndrome; data linkage; disease surveillance; linkage error; electronic health records; prevalence;

# Key messages

* Register and administrative data both indicated an increasing prevalence of Down’s syndrome among live births in England, but linked data suggest a stable prevalence.
* Analysis of Hospital Episode Statistics for England can be severely biased by linkage errors in the assignment of patient identifiers (‘HESID’) to hospital episode records, particularly when using birth episodes prior to 2009. Many administrative datasets can be similarly affected by errors in ‘internal linkage’.
* Linkage error is difficult to measure but quantitative bias analysis can be used to reflect plausible assumptions about its potential impact on an analysis.
* Linked data can provide more robust evidence for disease surveillance than single-source registry or administrative data and can support analysis involving changes in data collection systems.
* Linkage between datasets can be enhanced by identifying familial links within datasets, such as between mothers and children in Hospital Episode Statistics.

# Introduction

Congenital anomalies are a major cause of infant mortality, childhood morbidity and long-term disability, affecting over 1 in 50 children worldwide. (1) Accurate surveillance of congenital anomalies is essential to ensure that the right services are available to treat affected children, to provide reliable information on outcomes for prospective parents faced with difficult decisions in early pregnancy, to guide prevention programmes and for research into pregnancy and birth characteristics associated with anomalies. Congenital anomaly registries have been set up to collect accurate information for the surveillance of all anomalies. (2) Longitudinal population data routinely collected for administrative purposes (e.g. payments by health insurers and universal healthcare systems such as the NHS) offers an additional resource for identification of cases, information about long-term outcomes, and data on comparator populations to support analysis of aetiology and risk. However, the quality of administrative data is variable and its suitability for research applications requires careful evaluation. Linking data from independent sources of information about the same condition can be used to assess the quality of each source and to estimate the total number of cases, including those not detected by either source. Data linkage brings additional complexities, particularly around linkage error and integration of multiple sources of potentially conflicting information. In this article, we compare several possible approaches to analysis of linked population data, which we hope will provide methodological insights for population data science beyond their present application to Down's syndrome. (3)

The National Down's Syndrome Cytogenetic Register (NDSCR) was started in 1989 and collected all cytogenetic or DNA reports of trisomy 21 and the cytogenetic variants occurring in England and Wales. (4) In 2015 the NDSCR was incorporated into the National Congenital Anomaly and Rare Disease Registration Service (NCARDRS). (5) NCARDRS have expanded the systems available for follow-up of cases through administrative data and close links with notifiers in acute trusts. These changes in collection methods present a potential problem for research on trends during the transition period. It is important to have evidence that can help separate any effects of changes in data collection from changes in disease prevalence.

With England’s universal National Health Service (NHS), researchers and service planners alike are interested in establishing the potential for administrative NHS data to be used for population health monitoring and surveillance. Hospital Episode Statistics for England (HES) is a key source of hospital activity data (inpatient admissions, outpatient appointments and emergency department presentations) used for service planning and payment for hospital care funded by the National Health Service (NHS). (6) Linkage between registers has previously been used to assess case ascertainment (percent of cases detected) in the NDSCR and other congenital anomaly registers (4, 7), but not linkage to administrative data. Linkage between population-based registers and hospital records has been used to assess the coverage of each (8), but findings are specific to the data sources in questions.

In this article we describe the approach to linkage of NDSCR to HES and use the linked data to estimate the level of case ascertainment (proportion of all cases identified) in each data source and trends in the prevalence of Down's Syndrome among live births in England, between 1998 and 2013. In doing so, we aim to support integration of NDSCR and NCARDRS, to provide a resource for research on long-term outcomes of Down’s syndrome, and to establish methods that can be extended to linkage of NCARDRS and the full range of congenital anomaly and rare disease research that it facilitates.

# Methods

## Data sources: National Down Syndrome Cytogenetic Register

From 1989–2014, all cytogenetic laboratories in England and Wales notified the NDSCR of any cytogenetically confirmed diagnosis of trisomy 21 or related karyotype. (9) The register included pre- and postnatal diagnoses. Information on birth outcomes following prenatal diagnoses (live birth vs foetal death or termination) was obtained from clinicians and midwives but was missing in 8% of all diagnoses. (9) A total of 13 650 records were extracted for linkage, including 10 415 where the birth outcome was "live birth", the year of birth was between 1998 and 2013, and the postcode did not indicate residence outside of England (Figure S4, Appendix 1, Supplementary Material). A further 1226 records had missing birth outcomes but were within scope with respect to year and postcode region. The possible proportion of these that were live births is considered in the analysis, but these records had insufficient data for linkage.

## Data sources: Hospital Episode Statistics for England

With the universal healthcare provided by the NHS, HES captures 98–99% of all hospital activity in England. (6) When legally permitted and ethically justified, data are made available for research in de-personalised form (excluding names, addresses, etc.). Like most administrative data, records in HES represent events, in this case episodes of admitted patient care under one consultant, outpatient appointments, and emergency department presentations. Each patient’s records are linked through assignment of 'HESIDs' by NHS Digital, a process that is subject to linkage error. (10) Missed links lead to records belonging to the same patient being assigned different HESIDs, and false links (which are relatively rare) cause records from different patients to be assigned the same HESID. (11) The accuracy of HESIDs depends on the quality of matching data in the administrative record (NHS number, date of birth, postcode and local patient identifiers assigned by the treating hospital) which is known to be poorer in earlier years (10) and in birth episodes (12), partly because NHS numbers were not allocated at birth until after 2003.

Simply counting all distinct HESIDs with a Down’s syndrome diagnosis may lead to double-counting of cases when some patients have multiple HESIDs. To mitigate this, we first identified all birth episodes during our study period to identify a birth cohort which, by virtue of the fact that people are only born once, can reasonably be assumed to contain few duplicates. We then linked these to any subsequent hospital activity using HESID to identify diagnoses recorded after the birth admission. This birth cohort approach also allowed us to ensure that patients had been born in England and were therefore within the target population.

The birth cohort contained birth episodes for an estimated 10.3 million babies admitted during the 1997-98 to 2013-14 financial years. It therefore excludes babies born outside of hospital (2.2% of live births in England during 1997–2013) but represents 99.0% of the number of live births not at home recorded by the Office for National Statistics. (13) The construction of this birth cohort followed methods detailed in Harron, Gilbert (14) (Appendix 1, Supplementary Material). For each person in this cohort, Down's syndrome status was identified by the presence of any admitted patient care episode or outpatient appointment, at any time up until 2017-18 (linked by HESID), which included a three-character ICD10 "Q90" diagnosis code (i.e. including all four-character subclassifications).

While this birth cohort approach mitigates double-counting from the ‘splitting’ of patient's HESIDs, there is a potential trade-off in false negative misclassification when information about Down's Syndrome diagnoses is captured after birth but cannot be linked to the birth episode (because of having separate HESIDs). We therefore compared this approach to a simple analysis of HESIDs with dates of birth within the target range (but for whom country of birth could not necessarily be confirmed). These analyses were contrasted against analysis of NDSCR records alone, and to analysis of linked data from both NDSCR and HES.

## Data linkage

The NDSCR contains matching data relating to both the affected babies and their mothers (Table S1, Appendix 1, Supplementary Material). In HES, however, information about babies and mothers is recorded separately. Harron, Gilbert (14) demonstrated how babies can be linked to their mothers in pseudonymised HES (records with the personally identifying information usually required for linkage removed) using demographic and clinical variables captured in both admission records (mostly so-called 'baby tail' and 'maternity tail' variables). By extending their methods to the 1997-98–2013-14 financial years and incorporating additional matching data that were available at Public Health England, the birth cohort was enhanced by initial linkage of babies to their maternal delivery episodes within HES. Linkage of babies to mothers allowed maternal NHS numbers and dates of birth to be added to babies' HES records for linkage of HES to NDSCR, and for missing data in babies' postcodes to be completed using the mothers' records (Figure 1). Further details about HES cohort construction and probabilistic linkage  (15) are provided in Appendix 1 (Supplementary Material).

Figure 1 here

## Estimation of prevalence and case ascertainment

Three sets of estimates of annual prevalence were generated:

1. Using NDSCR data alone, with ONS estimates of live births in England (13) as the denominator. The main analysis considered only where the birth outcome was 'live birth'. An alternative analysis that included records with unknown birth outcomes is presented in Appendix 3 (Supplementary Material).
2. Using HES data alone. The main analysis considered the proportion of the HES birth cohort to have a Down's syndrome diagnosis code recorded at any time. This analysis includes only children born in NHS hospitals in England, in both numerator and denominator. Appendix 3 (Supplementary Material) presents an alternative analysis of the proportion of all HESIDs with dates of birth within the cohort window (i.e. without requiring an identified birth episode), ever to have a Down's syndrome diagnosis code recorded. That analysis includes children born outside hospital but also outside England, in both the numerator and denominator.
3. Using linked data and capture-recapture analysis to estimate the total number of incident live birth cases, with ONS estimates of live births in England (13) as the denominator (this analysis therefore includes children born in and outside hospital, in England, in both the numerator and denominator).

Prevalence estimation using linked data depends critically on the accuracy of linkage. If is the number of live birth diagnoses registered in NDSCR and is the number of people with Q90 diagnosis codes in the HES birth cohort, then the total number of incident cases, , can be divided into four key subgroups (Figure 2). Missed links between NDSCR and HES could result in somebody whose diagnosis is recorded in both sources being counted twice; once in and once in . False links could result in two people from and being counted once (in ). Capture-recapture analysis of the number of unrecorded cases () relies on accurate estimation of the other subgroups. (16) We therefore assigned estimates and plausible limits for both missed links and false links, varying the assigned rates of false links with the level of evidence supporting each link (Table S2, Appendix 2, Supplementary Material). Lastly, we also allowed for the possibility of false positives diagnostic codes to have been recorded in HES. Further details about the approach to quantitative bias analysis (17-19) and capture-recapture analysis are provided in Appendix 2 (Supplementary Material).

Figure 2 here

For each analysis, trends in prevalence were estimated using logistic regression of Down’s syndrome status on year of birth. This produced annual odds ratios that were converted into annual growth rates (% change in prevalence per year).

# Results

Results of linkage are included in Appendix 1 (Supplementary Material). Characteristics of candidate links and unlinked records are presented in Table 1 and stratified by match weight (a measure of agreement on variables used for linkage). Table S4 (Appendix 3, Supplementary Material) provides regional statistics.

Table 1 here

## Prevalence and trends

Analyses of single-source data are illustrated in Figures S8 and S9 (Appendix 3, Supplementary Material). A simple analysis of HESIDs (without restriction to a birth cohort) produced prevalence estimates that were both highest and most steeply increasing, with an estimated relative annual growth rate of 1.6% (95% CI: 1.3%, 2.0%), increasing to a prevalence of 13.1 cases per 10,000 live births in 2013. Restricting HES records to a birth cohort produced results that were comparable to analysis of NDSCR data if NDSCR records with unknown birth outcomes were included, with annual growth of 1.1% (95% CI: 0.6%, 1.5%) and 0.9% (95% CI: 0.5%, 1.3%), respectively. Excluding NDSCR records with unknown birth outcomes produced the lowest and most stable prevalence estimates, increasing by 0.4% (95% CI: 0.0%, 0.8%) per year to 10.4 cases per 10 000 live births in 2013.

In marked contrast to these indications of increasing prevalence, linked data indicated an overall prevalence that was generally higher than single-source estimates but was stable, with no significant change over time at a 95% confidence level (estimated annual growth = −0.1% (95% CI: −0.5%, 0.2%)) and an estimated prevalence of 12.3 cases per 10 000 live births, both in 2013 and overall (Figure 3).

Quantitative bias analysis provided regions of plausibility around these base cases estimates, reflecting uncertainty in the accuracy of diagnostic codes in HES and accuracy of linkage between HES and NDSCR (Figure 3 shading). Since 2006, prevalence estimates produced by analysis of the HES birth cohort have fallen within this range. Prevalence estimates produced by NDSCR live births were consistently below this range but would have overlapped it since 2004 if records with unknown birth outcomes were included.

Figure 3 here

Subtracting the number of cases captured in single-source estimates from the number estimated in linked data indicates that case ascertainment in NDSCR varied between 74% and 88% over the study period but was more stable than in HES, which increased from 81% (1998) to 96% (2012) (Figure 4).

Figure 4 here

# Discussion

This analysis demonstrates the feasibility and value of linking a perinatal cytogenetic register to HES. Data linkage provided a picture that contrasted with both the individual data sources which, if we accept it as being less biased, illustrates how linked data can be more than the sum of its parts.

These findings highlight strengths and limitations of both data sources. Ascertainment of live births with Down’s syndrome in the NDSCR appeared lower than in HES in more recent years, possibly because of loss to follow-up of prenatal diagnoses, but NDSCR's more consistent data quality over time provided better reflection of the underlying trends in prevalence. While HES appeared to have more complete case ascertainment in recent years, changes in the quality of HES over time could have created an alarming picture of an emerging epidemic of Down’s syndrome. We propose that this can largely be explained by decreasing errors in the assignment of HESIDs, which is evidenced by the increasing recording of NHS numbers in birth episodes up until about 2009 (Figure S7, Appendix 1, Supplementary Material) and the decreasing proportion of HES cases with only one episode in their first year (Figure S10, Appendix 3, Supplementary Material). This variation in quality of administrative data over time can distort analysis of trends and could confound evaluation of changes in policy or universal health care services, such as the recent introduction of free non-invasive screening for Down’s syndrome.

The problem of missed links in the assignment of HESIDs is poorly documented and there is little information available to help analysts address it. It is also statistically complex; in the simple analysis of all HESIDs is it likely to have contributed to overestimation from double-counting (‘splitting’ of one patient’s records into multiple observational units) while, in analysis of the birth cohort, it is likely to have contributed to underestimation through false negative misclassification of patients with missed links to their subsequent HES records that contain diagnosis codes. (17) Like many administrative datasets, HES is generated by service events that must be linked before person-level analyses can be implemented. With missed links in person identifiers having potential to wreak such statistical havoc, they are an issue that requires focused mitigation efforts by data providers through linkage quality assessment (20) and by researchers through sensitivity and bias analysis. (17-19)

With HES providing such a critical resource for health services research, and birth episodes containing unique information about perinatal health and family characteristics, there is a strong argument for refining the HESID algorithm to better support linkage of birth episodes by incorporating and leveraging familial links. While mother-child linkage is possible with the current data, a more complete record of familial links could be constructed with routine linkage to other health data, such as general practice registrations and routine child health checks.

Of course, the linked dataset is not a 'gold standard' and is itself prone to error. There are many possible sources of error and bias in this analysis, but we have attempted to quantify the main ones. People born earlier in the cohort had a longer period of observation in which HES events could be captured, but in the linked data this was accounted for by the capture-recapture analysis. More sophisticated methods for extracting diagnostic information from HES, using related diagnoses, procedure codes or outpatient appointments, may result in improved case ascertainment. Improvements in linkage might also be possible if NDSCR records with missing birth outcomes are considered, with approximate matching between date of sample and date of birth (sample date was not available for this linkage).

The main unaccounted potential source of error is in the assumptions of the capture-recapture analysis. Aside from no linkage error, the three main assumptions of the formula used are (i) equivalent source populations, (ii) homogeneity in probabilities of detection and (iii) independence of the data sources. (16) Most concerning is the source populations, which are known to have slight differences; people not born in hospital (an estimate 3% of all births) have no opportunity for their Down's status to be 'captured' in the HES birth cohort. By increasing the number of people identified only by NDSCR, this is likely to have led to overestimation of the number of unrecorded cases (cases detected by neither source), and therefore overestimation of the total prevalence and underestimation of case ascertainment in each dataset (the denominator for case ascertainment is the estimated total prevalence). We considered accounting for this quantitatively but this became complicated by implausible *combinations* with other bias parameters, suggesting that the potential for bias was already encompassed within the plausible range.

Heterogeneity in the probability of detection in either data source could have occurred if some parts of the population were both less likely to be screened and less likely to be born in hospital (e.g. people from rural and remote areas). This could similarly have inflated linked data estimates but may have been offset by dependence between the data source, if people recorded in one data source were more likely to be recorded in the other (e.g. because of related data collection mechanisms).

Regardless of these three assumptions, the contribution of unrecorded cases (those estimated through capture-recapture) was relatively small at between 0.8% (2012) and 6.2% (2001) (Table S3, Appendix 2, Supplementary Material). Even if these phenomena varied over time—which there is no obvious reason to suspect—they could not feasibly have accounted for all of the observed differences in trends.

Similarly, our findings with respect to identification of Down's syndrome in HES cannot be generalised to other diseases, phenotypes or datasets (see  (8) for a relevant example with contrasting findings). There is considerable variation in how diagnostic information is recorded in administrative data across diagnoses, between datasets and potentially over time, so it is important to assess the quality of administrative data sources in the context of each analysis.

When interpreting trends in live birth prevalence of Down's syndrome internationally, authors typically focus on the disentangling the competing effects of increasing risk factors for Down's syndrome (e.g. maternal age) and increasing rates of termination (21, 22). Using surveillance data, de Graaf, Buckley (23) estimate an increasing trend in live birth prevalence in the US over a similar time period. Using hospital records, the Public Health Agency of Canada (24) estimate a stable trend. Ours is the only known example to have used data linkage to control for quality issues in the underlying data sources.

# Conclusion

The differing conclusions that could be drawn from linked data versus single data sources highlight both the value that data linkage can offer and the dangers that it can pose when the quality of linkage is ignored. Reassuringly, we demonstrate that even when the quality of matching data is poor and there is uncertainty in linkage, quantitative bias analysis can be used to identify plausible boundaries within which target parameters should lie. (17) Probabilistic techniques (18) could further enhance quantification of this uncertainty.

When NDSCR and HES were linked, we found that detection of live birth cases in NDSCR increased over time, resulting in a slowly increasing trend in live birth prevalence of Down's syndrome. In HES, we observed a potentially alarming increase in prevalence that appeared partly attributable to internal linkage errors in the assignment of HESIDs. In the linked data, the trend appeared contrastingly stable. Given the value that this study demonstrates in linking registry with administrative data, the fairest basis for analysis of trends during and after transition from NDSCR to NCARDRS is likely to be provided by integrating NDSCR-HES linked data with a future linkage of NCARDRS to HES. Such routine linkage of registry and administrative data can provide other benefits also, in this case including invaluable support for analysis of long-term health outcomes.

# Acknowledgements

This linkage was implemented at Public Health England with support from the National Congenital Anomalies and Rare Disease Registration Service (NCARDRS). Linkage between NDSCR and HES was conducted to facilitate the research project ‘Evaluating variation in special educational needs provision for children with Down syndrome and associations with emergency use of hospital care’, with approval from the Health Research Authority’s London – Camden and King’s Cross Research Ethics Committee (ref: 16/LO/0094) and Confidentiality Advisory Group (ref: 16/CAG/0015), and the Administrative Data Research Network (ref: PROJ-165). We thank Sarah Stevens, Helen Duncan and Julian Flowers for facilitating support for this project by PHE. We acknowledge with thanks all the clinical, administrative and laboratory staff who routinely supply NCARDRS with information, without whom this data would not be available for research. We acknowledge advice provided by the NCARDRS team concerning data collection practices and problems that could influence this analysis, and comments provided on drafts of this paper by members of the NCARDRS team.

This work was supported by the Economic and Social Research Council [grant number: ES/L007517/1], the Medical Research Council [grant number: London MR/ K006584/1], the NIHR Great Ormond Street Hospital Biomedical Research Centre, Health Data Research UK [RG] and The Wellcome Trust (grant number: 103975/A/14/Z; KH).

# References

1. Boyle B, Addor M-C, Arriola L, et al. Estimating Global Burden of Disease due to congenital anomaly: an analysis of European data. *Archives of Disease in Childhood - Fetal and Neonatal Edition* 2018; **103**: F22-F8.

2. Boyd PA, Haeusler M, Barisic I, Loane M, Garne E, Dolk H. Paper 1: The EUROCAT network—organization and processes†. *Birth Defects Research Part A: Clinical and Molecular Teratology* 2011; **91**: S2-S15.

3. McGrail K, Jones K, Akbari A, et al. A Position Statement on Population Data Science: The Science of Data about People. *IJPDS* 2018; **3**.

4. Savva GM, Morris JK. Ascertainment and accuracy of Down syndrome cases reported in congenital anomaly registers in England and Wales. *Archives of Disease in Childhood - Fetal and Neonatal Edition* 2009; **94**: F23-F7.

5. Stevens S, Miller N, Rashbass J. Development and progress of the National Congenital Anomaly and Rare Disease Registration Service. *Arch Dis Child* 2018; **103**: 215-7.

6. Herbert A, Wijlaars L, Zylbersztejn A, Cromwell D, Hardelid P. Data Resource Profile: Hospital Episode Statistics Admitted Patient Care (HES APC). *Int J Epidemiol* 2017; **46**: 1093-i.

7. Morris JK, Grinsted M, Springett AL. Accuracy of reporting abortions with Down syndrome in England and Wales: a data linkage study. *Journal of Public Health* 2016; **38**: 170-4.

8. Bourke J, Wong K, Leonard H. Validation of intellectual disability coding through hospital morbidity records using an intellectual disability population-based database in Western Australia. *BMJ open* 2018; **8**: e019113-e.

9. Morris JK, Springett A. *The National Down Syndrome Cytogenetic Register for England and Wales: 2013 Annual Report*: Queen Mary University of London, Barts and The London School of Medicine and Dentistry; 2014.

10. Health and Social Care Information Centre. *Replacement of the HES Patient ID (HESID)*; 2009.

11. Hagger-Johnson G, Harron K, Gonzalez-Izquierdo A, et al. Identifying Possible False Matches in Anonymized Hospital Administrative Data without Patient Identifiers. *Health Serv Res* 2015; **50**: 1162-78.

12. Hagger-Johnson G, Harron K, Fleming T, et al. Data linkage errors in hospital administrative data when applying a pseudonymisation algorithm to paediatric intensive care records. *BMJ Open* 2015; **5**: e008118.

13. Office for National Statistics. *Number of live births at home and total live births, England, 1994 to 2014 birth registrations*; 2016.

14. Harron K, Gilbert R, Cromwell D, van der Meulen J. Linking Data for Mothers and Babies in De-Identified Electronic Health Data. *PLoS One* 2016; **11**: e0164667.

15. Doidge JC, Harron K. Demystifying probabilistic linkage: Common myths and misconceptions. *IJPDS* 2018; **3**.

16. Stephen C. Capture-Recapture Methods in Epidemiological Studies. *Infect Control Hosp Epidemiol* 1996; **17**: 262-6.

17. Doidge J, Harron K. Linkage error bias. *Int J Epidemiol* 2019; **(in press)**.

18. Lash TL, Fox MP, Fink AK. *Applying Quantitative Bias Analysis to Epidemiologic Data*. New York: Springer; 2009.

19. Lash TL, Fox MP, MacLehose RF, Maldonado G, McCandless LC, Greenland S. Good practices for quantitative bias analysis. *Int J Epidemiol* 2014: 1969-85.

20. Harron KL, Doidge JC, Knight HE, et al. A guide to evaluating linkage quality for the analysis of linked data. *Int J Epidemiol* 2017; **46**: 1699-710.

21. de Graaf G, Buckley F, Skotko BG. Estimates of the live births, natural losses, and elective terminations with Down syndrome in the United States. *American Journal of Medical Genetics Part A* 2015; **167**: 756-67.

22. Morris JK, Mutton DE, Alberman E. Revised estimates of the maternal age specific live birth prevalence of Down's syndrome. *J Med Screen* 2002; **9**: 2-6.

23. de Graaf G, Buckley F, Dever J, Skotko BG. Estimation of live birth and population prevalence of Down syndrome in nine U.S. states. *American Journal of Medical Genetics Part A* 2017; **173**: 2710-9.

24. Public Health Agency of Canada. *Down Syndrome Surveillance in Canda, 2005-2013*: Public Health Agency of Canada; 2017.

# Abbreviations

HES: Hospital Episode Statistics for England (NHS Digital)

HESID: A unique patient pseudo-identifier assigned by NHS Digital to identify hospital episode records that relate to the same person

NCARDRS: National Congenital Anomaly and Rare Disease Registration Service (Public Health England)

NDSCR: National Down Syndrome Cytogenetic Register (Public Health England)

NHS: National Health Service

NHS Digital: Trading name of the Health and Social Care Information Centre

Table 1: Characteristics of linked and unlinked records

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Deterministic links | Probabilistic (MW > 40.6) | Probabilistic (MW: 30.5–40.6) | Probabilistic (MW: 18.1–30.5) | Probabilistic (MW < 18.1) | Unlinked NDSCR records | Unlinked HES cases |
|   |  |  |  |  |  |  |  |
| *NDSCR records* | 4939 | 3694 | 449 | 662 | 534 | 137 | ─ |
| *HES records* | 4941 | 3703 | 446 | 646 | 654 | ─ | 2280 |
| *Candidate links1*  | 4941 | 3720 | 454 | 799 | 739 | ─ | ─ |
| Q90 code (in HES records) | 96.4% | 91.1% | 70.2% | 81.4% | 17.0% | ─ | ─ |
| Difference in DOB > 180 days (in candidate links) | 0.4% | < 0.3% | 1.5% | 3.6% | 6.0% | ─ | ─ |
| Sex = male |  |  |  |  |  |  |  |
| *in NDSCR records* | 55.4% | 53.9% | 52.2% | 52.9% | 54.7% | 49.3% | ─ |
| *in HES records* | 55.1% | 53.8% | 52.7% | 52.9% | 56.3% | ─ | 53.6% |
| Premature (<37 weeks) |  |  |  |  |  |  |  |
| *in NDSCR records* | 22.3% | 19.1% | 16.7% | 12.7% | 18.4% | 10.5% | ─ |
| *in HES records* | 23.3% | 22.3% | 22.5% | 20.7% | 10.8% | ─ | 23.3% |
| Age at diagnosis (in NDSCR records) |  |  |  |  |  |  |  |
| *Prenatal* | 9.9% | 10.0% | 7.1% | 3.7% | 8.5% | 7.4% | ─ |
| *< 12 months* | 89.5% | 89.7% | 91.9% | 93.7% | 81.4% | 85.2% | ─ |
| *≥ 12 months* | 0.6% | 0.3% | 1.0% | 2.6% | 10.1% | 7.4% | ─ |
| Age at first diagnosis code (in HES records) |  |  |  |  |  |  |  |
| *< 12 months* | 90.9% | 89.8% | 90.4% | 88.2% | 88.9% | ─ | 77.7% |
| *≥ 12 months* | 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% |
| DOB: Date of birth; HES: Hospital Episode Statistics for England; MW: match weight; NDSCR: National Down Syndrome Cytogenetic Register.NDSCR records exclude those with missing birth outcome. All data are column proportions, ignoring missing data, so that associations between record characteristics and linkage quality are reflected by differences in proportion across columns within each row. Probabilistic links are grouped by 'match weight', a score reflecting the level of agreement over matching variables (see Methods). 1The number of candidate links may be higher than the number of records in either file, indicating ambiguity of multiple links with equal agreement; for two of such candidate links, either at least one is false or both are true and it is the records in the contributing files that have not been completely deduplicated.Source: Hospital Episode Statistics (HES), NHS Digital (Copyright © 2019. Re-used with the permission of NHS Digital. All rights reserved) and the National Down Syndrome Cytogenetic Register (NDSCR), Public Health England. |

Hospital episodes

Maternal cohort

Birth cohort

Enhanced birth cohort

NDSCR registrations

Down's syndrome cases

(i)

(ii)

(iii)

Figure 1: Linkage overview

1. Construction of cohorts of live babies born and mothers who delivered a live baby in HES
2. Linkage of babies to mothers in HES
3. Linkage of the enhanced HES birth cohort to NDSCR

Figure 2: Subgroups for estimating prevalence and case ascertainment

 Number of cases

 Number of cases who appear as registered live birth diagnoses in NDSCR

 Number of cases who appear as in the HES birth cohort and have an associted “Q90” diagnosis code for Down’s syndrome recorded at any time in HES.

 Number of cases, with presence in NDSCR () and HES () indicated by 1 (present) or 0 (absent), such that indicates the number of unrecorded cases.

Figure 3: Annual prevalence of Down's syndrome in England, 1998–2013: Comparison of three data sources

NDSCR estimates are derived using ONS births data as denominator; HES estimate is derived from the HES birth cohort data only so exclude children born outside hospital; Linked data estimates are derived using capture-recapture methods with ONS births data as denominator and quantitative bias analysis to generate plausible limits of uncertainty. Trend lines indicate fitted linear regressions of annual prevalence estimates, with slopes representing the estimated annual change in prevalence over the study period.
Source: Hospital Episode Statistics (HES), NHS Digital (Copyright © 2019. Re-used with the permission of NHS Digital. All rights reserved) and the National Down Syndrome Cytogenetic Register (NDSCR), Public Health England.

Figure 4: Ascertainment of Down's syndrome in the National Down Syndrome Cytogenetic Register and Hospital Episode Statistics for England

Proportions are the observed number of incident cases in each dataset (live births in NDSCR and birth cohort cases with Q90 diagnosis codes at any time in HES) divided by the base case estimated number of incident cases in England using capture-recapture analysis of linked data.
Source: Hospital Episode Statistics (HES), NHS Digital (Copyright © 2019. Re-used with the permission of NHS Digital. All rights reserved) and the National Down Syndrome Cytogenetic Register (NDSCR), Public Health England.

Appendix 1: Methods for cohort construction and linkage

## Construction of cohorts of babies and mothers in Hospital Episode Statistics for England

The unit of recording in the Admitted Patient Care section of HES is an 'episode' of care under one consultant. An admission may be comprised of multiple episodes and patients may have multiple admissions over time, plus attendances in outpatient clinics and emergency departments. These are linked by NHS Digital, who assign each record a 'HESID' indicating a distinct patient. Being a linkage procedure, allocation of HESIDs is subject to linkage error; missed links that result in people's records being allocated different HESIDs and false links that result in different people sharing one HESID. There has been little evaluation of the algorithm used to assign HESIDs but previous experience highlighted increased error rates in birth episodes, stemming from the allocation of NHS numbers after birth registration (i.e. *after* discharge from birth admissions) and a known error in recording of infants' postcodes prior to 2011 (1, 2).

To mitigate errors in HESID, we adopted methods for combining episodes relating to the same person that did not rely solely on HESID, based on those described in Harron, Gilbert (1). On extending these methods to the 1997-98 to 2001-02 years, changes in the way that baby/maternity tail variables were recorded during the earlier years meant that additional criteria had to be incorporated. The processes for constructing the birth cohort is summarised in Figure S1.

11 548 606 episodes meeting inclusion criteriaa

11 440 863 episodes

107 743 unfinished episodes and stillbirths excluded

11 262 158 records

178 705 episodes merged using original criteriab

11 245 468 records

16 690 duplicates with different HESIDs excludedc

10 326 665 estimated
live births

918 803 records merged using additional criteriad

Figure S1 Construction of the HES birth cohort

a As per Appendix S1 of Harron, Gilbert (1). b As per Harron, Gilbert (1):Match on HESID, start date, age, postcode, birth order and birth weight.
c As per Harron, Gilbert (1): Match on start date, age, hospital, GP practice, ethnicity, date of birth, and baby tail field but with different HESIDs.
d No more than two out of 21 potential inconsistencies on demographic and baby tail variables within HESID.
Source: Hospital Episode Statistics (HES), NHS Digital (Copyright © 2019. Re-used with the permission of NHS Digital. All rights reserved)

Similarly, additional criteria had to be incorporated into the processing of maternal records (Figure S2). This reflected both changes in the way maternity tail variables were recorded over time and the additional requirement in this application to link data for multiple births. Each maternal admission record that indicated a multiple birth was reshaped to create a separate record for each baby, and additional exclusion rules then applied to address the large volumes of invalid/not applicable codes contained in the multiple birth fields (the maternity tail variables indexed by "\_[N]") required additional exclusion criteria to be applied to these reshaped maternal records (Figure S2, footnote (d)).

12 654 602 episodes meeting inclusion criteriaa

12 463 972 episodes

190 630 episodes excluded by original criteriab

10 273 940 episodes

2 190 032 episodes excluded by additional criteriac

10 069 758 records

204 182 episodes < 169 days apart merged

12 004 082 records after reshaping multiple births

10 262 663 maternal records of babies born

1 741 419 records excluded by additional criteriad

Figure S2 Construction of the maternal cohort

a As per Appendix S1 of Harron, Gilbert (1).
b As per Appendix S1 of Harron, Gilbert (1).
c Less than 2 valid maternity codes after cleaning *and*no relevant procedure code *and* no relevant diagnostic code in the first five diagnosis fields.
d Birth characteristics relating to different babies in multiple births are indexed by \_[N] (e.g. birthweight\_1 birthweight\_2, etc) which becomes a variable when these data are reshaped. The field 'NUMBABY' separately records the number of babies associated with a delivery. Reshaped records were excluded whenever: (index > 3) or (index = 3 and NUMBABY = 1) or (index = 2 or 3, and the number of babies indicated by reshaping was greater than 5, and NUMBABY = 1 or more than 5).
Source: Hospital Episode Statistics (HES), NHS Digital (Copyright © 2019. Re-used with the permission of NHS Digital. All rights reserved)

Comparison of the cohort sizes to estimates of live births in hospitals generated from birth registration records by the Office for National Statistics suggest that any double-counting arising from the allocation of multiple HESIDs to single patients within the birth cohort is minimal (Figure S3).

Figure S3 Comparison of cohort sizes to ONS estimates of births in hospital in England

External reference data was births not at home, derived from birth registration data by  1ONS data are calendar years; HES data are financial years commencing, so some difference is expected
Source: Hospital Episode Statistics (HES), NHS Digital (Copyright © 2019. Re-used with the permission of NHS Digital. All rights reserved) and the Office for National Statistics (3).

## Selection of records from the National Down Syndrome Cytogenetic Register

13 663 NDSCR records from 1996–2013 extracted1

13650 NDSCR records submitted for linkage:

* 12 255 live births
* 1395 with missing birth outcome

13 duplicates merged

10 415 NDSCR records of live birth within scope

2009 records excluded from analysis, including:

* 641 non-English postcodes2
* 1443 estimated year of birth outside 1998–20132

1226 records missing birth outcome but otherwise within scope

Figure S4 Selection of records from the National Down Syndrome Cytogenetic Register (NDSCR)

1 Extraction and linkage initially included records with year of birth = 1997 or year of sample = 1996 for prenatal diagnoses. Because HES records for 1997 were not available for the full calendar year, this analysis excludes these records.
2 Including 75 records with both criteria

## Linkage of babies to mothers in Hospital Episode Statistics for England

The linkage implemented by Harron, Gilbert (1) involved 23 'pseudonymised' matching variables; mostly clinical fields contained in the baby and maternity tails, plus postcode district (derived by NHS Digital from HOMEADD), mother's age (MATAGE, derived by NHS Digital from the mother's and baby's dates of birth) and an estimated date of birth for the baby (derived by the authors from the date of procedure or admission). Our linkage was additionally supported by access to full postcodes and dates of birth. Postcodes for babies, if missing, were imputed (carried backwards) from the first non-missing admission or outpatient appointment for that HESID.

As with Harron, Gilbert (1) we used a two-step linkage procedure involving an initial deterministic step using a subset of matching variables that uniquely identify some individuals, and a probabilistic step that used all matching variables. The deterministic step provided a reference set for estimation of values (the probability that a matching variable agrees if the records are a match) for use in calculation of match weights in the probabilistic step. Our deterministic step combined two rules:

1. Unique agreement on financial year, hospital trust, general practice, sex, birth order, gestational age and mother's age, with no disagreement on infant's date of birth or mother's date of birth, *or*
2. Unique agreement on financial year, hospital trust, infant's date of birth and mother's date of birth, with no disagreement on general practice, sex, birth order or mother's age

Use of variables in the deterministic step precludes estimating their values using this data. For variables used in deterministic linkage, values were informed by previous implementation (1). values for the remaining matching variables were estimated as the proportion of deterministic links exhibiting agreement on each matching variable. values were estimated using random draws and were value-specific where possible. For each pair of records, partial match weights were summed across all matching variables to calculate an overall match weight, assigning partial weights of zero in the presence of missing values.

The probabilistic step only considered records that matched on hospital trust and in which the baby's admission commenced no more than seven days before the mother's admission commenced and no more than seven days after the mother's admission ended. Candidate links were ranked by match weight and sorted into 'unambiguous links' in which the top-ranked infant record for a maternal record and the top-ranked maternal record for an infant record were consistent and uncontended, 'multiple links' in which the top rank was shared by multiple candidates (which could all be true, given potential linkage errors in the construction of the cohorts) and 'ambiguous links' in which the highest ranks were inconsistent. An iterative sample of record pairs was clerically reviewed to select a minimum threshold for accepting these candidate links (in this case, match weight ≥ 3). In summary, the probabilistic linkage steps were to:

1. Identify mother and baby records from the same hospital that are no more than seven days apart in time.
2. Calculate match weights and rank all candidate links by match weight
3. Use iterative, sampled clerical review to decide a minimum match weight for accepting links.
4. Accept unambiguous links above this threshold, where the highest ranked baby for a mother is the same as the highest ranked mother for that baby.
5. Flag links above the threshold where there is inconsistency the highest ranked pairs as potential errors in linkage.
6. Flag links above the threshold where there is ambiguity in the highest ranked pairs as potential errors in linkage or true multiple links (reflecting multiple records for the same mother or baby in the data).

Overall, 49.7% of infants in the birth cohort were able to be deterministically linked to a mother and, for a further 44.0%, an unambiguous probabilistic link was identified. Small numbers of multiple (0.5%) and ambiguous (1.5%) links were found, with 4.3% of births remaining unlinked (Figure S5). Linkage of baby to mother was slightly less likely if the HES record indicated Down's syndrome (88.9% of birth cohort members with any Q90 diagnosis codes were able to be linked to a maternal record, compared with 95.7% of birth cohort members without Q90 diagnosis codes). Babies with no linked maternal record had fewer variables on which to link with NDSCR.

Figure S5 Linkage of babies to mothers in Hospital Episode Statistics for England, by financial year

Source: Hospital Episode Statistics (HES), NHS Digital (Copyright © 2019. Re-used with the permission of NHS Digital. All rights reserved).

## Linkage of enhanced HES birth cohort to NDSCR

After enhancing the birth cohort with matching variables from their linked mothers, potential matching variables available for the HES-NDSCR linkage included: NHS numbers for both mother and child, dates of birth for both mother and child (each split into day, month and year to accommodate partial entries in NDSCR), postcode (split into two parts to accommodate partial entries), sex, gestational age, birth weight, multiple birth status, and Down's Syndrome status (constant in NDSCR, and as indicated by diagnosis codes in HES).

Linkage involved an initial deterministic (rule-based) step, which supported a subsequent probabilistic step, involving 'match weights' (scores, based on the Fellegi and Sunter (4) framework). The deterministic linkage used child's NHS number with clerical review of all returned links that disagreed on other matching variables. The identified links were then used to estimate values (the proportions of true links that agree on each matching variable) for the remaining matching variables, and values (the proportions of true non-links that agree on each matching variable) were estimated using a random draw. These and values were used to construct match weights, that were used to rank candidate HES links for each NDSCR record, to identify the most likely candidates. The highest ranking candidate HES record for each NDSCR record was retained and stratified according to match weight, indicating the degree of correspondence between the records. Estimated and values and their corresponding partial match weights are provided in Table S1, and completeness of matching variables is illustrated in Figure S6 and Figure S7.

Table S1 Match weights in HES-NDSCR linkage

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Matching variable | 1 | 2 | Partial weightif agree3 | Partial weightif disagree4 |
| Day of birth | 0.99 | 0.03 | 4.92 | -6.09 |
| Month of Birth | 0.99 | 0.08 | 3.58 | -7.50 |
| Year of Birth | 1.00 | 0.06 | 4.17 | -8.28 |
| Sex | 0.99 | 0.50 | 0.98 | -5.86 |
| Birth weight | 0.90 | 2.89E-03 | 8.27 | -3.26 |
| Gestational age | 0.77 | 0.14 | 2.49 | -1.91 |
| Multiple birth flag | 0.99 | 0.92 | 0.11 | -2.86 |
| Down's syndrome status | 0.96 | 1.16E-03 | 9.69 | -4.79 |
| Postcode (first part) | 0.95 | 7.30E-04 | 10.35 | -4.43 |
| Postcode (second part) | 0.89 | 3.17E-04 | 11.46 | -3.22 |
| Mother's NHS number | 0.96 | 1.00E-06 | 19.87 | -4.58 |
| Mother's day of birth | 0.97 | 0.03 | 4.90 | -5.23 |
| Mother's month of birth | 0.97 | 0.08 | 3.54 | -5.13 |
| Mother's year of birth | 0.98 | 0.03 | 4.97 | -5.65 |
| 1Estimated proportion of (true) matches exhibiting agreement on matching variable, if not missing, derived from observed proportion among deterministic links.2Estimated proportion of (true) non-matches exhibiting agreement on matching variable, derived from random sample of all record pairs.3 4  |

Figure S6 Completeness of NDSCR matching variables

Source: National Down Syndrome Cytogenetic Register (NDSCR), Public Health England.

Figure S7 Completeness of HES matching variables (for HES-NDSCR linkage)

Child's date of birth, and Down's Syndrome status (as indicated by diagnosis codes) were complete in all years
Source: Hospital Episode Statistics (HES), NHS Digital (Copyright © 2019. Re-used with the permission of NHS Digital. All rights reserved).

Because NHS numbers were largely missing in both HES and NDSCR prior to 2003 (Figures S5 and S6, Supplementary Appendix 1), linkage between NDSCR and HES relied mostly on probabilistic techniques in these years. NHS numbers were also entirely missing in the prenatal NDSCR records with missing birth outcome. Of the NDSCR records of live births that were within scope, 4939 (47.4%) were linked deterministically to members of the HES birth cohort, 96.4% of whom also had Q90 diagnosis codes (see Table 1 in main article). The existence of two deterministic links for each of two NDSCR records indicated a very small degree of residual double-counting in the HES birth cohort (instances where the same person was probably represented twice; 0.04% of deterministically linked records, but potentially higher among others). Probabilistic linkage identified candidate links for another 5339 (51.3%) live birth NDSCR records.

Of the deterministic links and probabilistic links with match weights greater than 18.1, most had Down's syndrome diagnosis codes. Of the probabilistic links with match weights below 18.1, few had diagnosis codes but this is to be expected given the contribution of diagnosis codes towards match weights (Table S1). Female records were slightly less likely to be linked, as were postnatal diagnoses, especially those occurring after 12 months of age. HES cases with multiple admission episodes in their first year of life were also more likely to be linked to NDSCR than those with only a single episode.

# Appendix 1 References

1. Harron K, Gilbert R, Cromwell D, van der Meulen J. Linking Data for Mothers and Babies in De-Identified Electronic Health Data. *PLoS One* 2016; **11**: e0164667.

2. Hagger-Johnson G, Harron K, Fleming T, et al. Data linkage errors in hospital administrative data when applying a pseudonymisation algorithm to paediatric intensive care records. *BMJ Open* 2015; **5**: e008118.

3. Office for National Statistics. *Number of live births at home and total live births, England, 1994 to 2014 birth registrations*; 2016.

4. Fellegi I, Sunter A. A theory for record linkage. *J Am Stat Assoc* 1969; **64**: 51-79.

Appendix 2: Methods for quantitative bias analysis and capture-recapture analysis

This appendix provides additional explanation of the quantitative bias analysis and capture-recapture analysis. Bias parameters (rates of different types of error and their plausible limits) were estimated through author consensus, then combined with capture-recapture methods (1) to estimate the number of unrecorded cases and total incident live births with Down’s syndrome in England. The assigned bias parameters are summarised in Table S2 and stepped results are summarised in Table S3.

## Potential false positive diagnostic codes in HES

Because it was derived from cytogenetic laboratories, we assumed that NDSCR records would not include false positive *diagnoses* (records with unknown birth outcomes may not all have been liveborn, but these were excluded from analysis). For HES, we evaluated the positive predictive value of Q90 diagnosis codes (PPV; the proportion of records with codes that truly have Down's syndrome) by examining the proportion of cases where only a single code was recorded despite many records existing for that HESID. For all cohort HESIDs that had a Q90 diagnosis code in any record, there were a total of 132,855 admitted patient care episodes. Of these, 95,593 (70.5%) included a Q90 code. By restricting this to HESIDs with at least 5 or 10 episodes (7784 and 4316 HESIDs, respectively), we could see that 95.3% (7419) and 95.5% (4121) had at least two episodes containing a diagnosis code. It seems reasonable to assume that the PPV for *multiple* Q90 codes is approximately 100%. These statistics therefore support a minimum plausible limit of 95.0% for the PPV of having *any* Q90 code and a true PPV that is likely higher. For the analysis of linked data, we assigned a base case PPV of 99.5% to having *any* Q90 code, with plausible limits of 95.0–100.0%

## Estimates of linkage error

Because all deterministic link were based on unique identifiers and all deterministic links with high levels of disagreement on other matching variables had been clerically reviewed (with any questionable links subjected to probabilistic linkage instead), we assumed that the precision of linkage (proportion of links that are true) was 100% for these. For probabilistic links, we assigned point estimates and plausible limits for precision that decreased with match weight from 99–100% above a match weight of 40.6, down to between 50–100% at match weights between 0.0 and 18.1 (Table S). For the proportion of unlinked records that were in truth missed links (cases that truly appear in both datasets but for which a link could not be identified), we had little to base estimates on so assigned wide plausible limits of 10–90%. For example, if there were 50 unlinked NDSCR records and 100 unlinked HES cases in a given year, then the maximum possible number of links between these was 50 and we estimated that between 5 (10%) and 45 (90%) of these were missed links.

## Capture-recapture analysis

From the basic formula for capture-recapture analysis, the number of unrecorded cases can be given by the formula  (1). Other than an absence of linkage error, there are three further assumptions of this formula: (i) that the data sources are independent, (ii) that cases are homogenous (equal) in terms of their probability of detection and (iii) that the population sampled by each data source is identical. In applications of capture-recapture to disease surveillance, these assumptions are often not met (2). We therefore liaised with data collectors to qualitatively assess each assumption’s plausibility and potential implications (see Discussion).

## Formulae for quantitative bias analysis and capture-recapture analysis

1. Analysis of linked data without accounting for any potential sources of error
For each year, calculate the number of cases as:
2. Adjust for **false positive diagnoses** in HES

For each scenario (upper, base case, lower) and each year, estimate the total number of cases correctly positively identified by Q90 diagnosis codes in the HES birth cohort as:

where is the positive predictive value for Down’s syndrome status given the presence of at least one Q90 diagnosis code at any time (a defined bias parameter).

1. Estimate number of true matches (record pairs that should link), adjusting for **false links**
For each category of links (deterministic plus four categories of probabilistic), each scenario, and each year, calculate:

where is the lower of the number of NDSCR records or HES records that contribute to a set of identified candidate links (i.e. assume one-to-one linkage, so that for 10 records from file A and 20 records from file B that form a set of candidate links, there can be at most 10 links in that set), and is the proportion of links that are true matches (the positive predictive value of linkage).

1. Estimate number of true matches (record pairs that should link), adjusting for false links and **missed links**

For each scenario and each year, calculate:

where is the estimated proportion of unlinked records that are missed links (a defined bias parameter, relating to the sensitivity or recall of linkage) and the number of unlinked records is taken from the lower of the estimated number of unlinked NDSCR records or unlinked HES records with diagnosis codes, after combining the estimated number of false links from Step 2 with the number of each that have no candidate links in each dataset.

1. Estimate number of true Q90 diagnoses in HES that have true matches in NDSCR, adjusting for **false links**
For each scenario and each year, repeat Step 2 using the estimated subset of HES records that have true Q90 diagnosis codes from Step 1 (for simplicity, in this step we assumed that false positive diagnoses were concentrated among the unlinked HES records with diagnosis codes, so that HES records with both diagnosis codes and links were assumed to be true Q90 diagnoses but not necessarily true links).
2. Estimate number of true Q90 diagnoses in HES that have true matches in NDSCR adjusting for false links and **missed links** ()

For each scenario and each year, repeat Step 3 using the estimated number of matches from Step 4, the estimated number of unlinked Q90 diagnoses in HES implied by Step 4, and the same estimated number of unlinked NDSCR records implied by Step 2 and used in Step 3.

1. Estimate total cases, including **unrecorded cases**
For each year (and each of the scenarios produced above) calculate:
where is the number of live birth diagnoses registered in NDSCR, is the estimated number of live births with true Q90 diagnoses in the HES birth cohort (from Step 1), and is the estimated number true matches between these (from Step 5).
	1. Unrecorded cases can now be derived as:
	2. And case ascertainment can now be derived as and

Table S2 Bias parameter estimates used in analysis of linked data

| Bias parameter | Analysisa |
| --- | --- |
| Lower limit | Base case | Upper limit |
| **Linkage precision, by link quality** |  |  |  |
| Deterministic | 100.0% | 100.0% | 100.0% |
| Probabilistic (match weight > 40.6) | 100.0% | 100.0% | 99.0% |
| Probabilistic (match weight: 30.5-40.6) | 100.0% | 98.0% | 95.0% |
| Probabilistic (match weight: 18.1-30.5) | 100.0% | 90.0% | 80.0% |
| Probabilistic (match weight < 18.1) | 100.0% | 80.0% | 50.0% |
| **Proportion of unlinked records that are missed links** | 90.0% | 50.0% | 10.0% |
| **Positive predictive value of diagnosis codes among unlinked HES cases** | 95.0% | 99.5% | 100.0% |
| aEstimates and plausible limits assigned by author consensus (see text for further explanation). Limits are arranged such that lower limits of each parameter translate into the lowest estimates of prevalence using linked data. |

Table S3 Estimated incidence of Down's Syndome, sequentially adjusted for each source of possible error

| Year | Estimated number of cases, by analysis step |
| --- | --- |
| Live birth diagnoses in NDSCR | Live births with Q90 codes in HES birth cohort | Accepting maximum number of candidate links | Adjusted for false positive diagnoses a | …and adjusted for false links a | …and adjusted for missed links a | …and adjusted for undetected cases a |
| 1998 | 598 | 562 | 696 | 693 | 701 | 684 | 707 |
| 1999 | 571 | 583 | 689 | 686 | 705 | 681 | 705 |
| 2000 | 579 | 547 | 684 | 681 | 702 | 674 | 702 |
| 2001 | 550 | 565 | 712 | 709 | 741 | 695 | 741 |
| 2002 | 573 | 583 | 707 | 704 | 727 | 696 | 727 |
| 2003 | 584 | 640 | 716 | 713 | 728 | 710 | 728 |
| 2004 | 629 | 687 | 757 | 754 | 761 | 748 | 761 |
| 2005 | 700 | 717 | 793 | 789 | 794 | 784 | 794 |
| 2006 | 709 | 745 | 819 | 815 | 816 | 806 | 816 |
| 2007 | 676 | 707 | 774 | 770 | 770 | 762 | 770 |
| 2008 | 694 | 727 | 790 | 786 | 790 | 782 | 790 |
| 2009 | 730 | 776 | 837 | 833 | 833 | 825 | 833 |
| 2010 | 689 | 742 | 779 | 775 | 776 | 772 | 776 |
| 2011 | 702 | 778 | 825 | 821 | 827 | 819 | 827 |
| 2012 | 740 | 831 | 880 | 876 | 873 | 866 | 873 |
| 2013 | 691 | 761 | 819 | 815 | 819 | 810 | 819 |
| a Base case estimates, rounded to nearest integer |

# References

1. Stephen C. Capture-Recapture Methods in Epidemiological Studies. *Infect Control Hosp Epidemiol* 1996; **17**: 262-6.

2. Braeye T, Verheagen J, Mignon A, et al. Capture-Recapture Estimators in Epidemiology with Applications to Pertussis and Pneumococcal Invasive Disease Surveillance. *PLoS One* 2016; **11**: e0159832.

Appendix 3: Supplementary results

Figure S8 Annual number of Down's Syndrome cases detected in separate data sources

HES: Hospital Episode Statistics for England; NDSCR: National Down Syndrome Cytogenetic Register
Source: Hospital Episode Statistics (HES), NHS Digital (Copyright © 2019. Re-used with the permission of NHS Digital. All rights reserved) and the National Down Syndrome Cytogenetic Register (NDSCR), Public Health England.

Figure S9 Annual prevalence of Down's Syndrome in separate data sources

HES: Hospital Episode Statistics for England; NDSCR: National Down Syndrome Cytogenetic Register. The denominator in HES is the estimated number of births in the birth cohort or number of HESIDs in the whole of HES, for each year of birth; the denominator for NDSCR is the estimated number of live births in England reported by the Office for National Statistics (1) (see *Estimation of prevalence and case ascertainment* for explanation)
Source: Hospital Episode Statistics (HES), NHS Digital (Copyright © 2019. Re-used with the permission of NHS Digital. All rights reserved) and the National Down Syndrome Cytogenetic Register (NDSCR), Public Health England.

Table S4 Geographic regions of linked and unlinked records

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|   | Deterministic | Probabilistic (MW > 40.6) | Probabilistic (MW: 30.5–40.6) | Probabilistic (MW: 18.1–30.5) | Probabilistic (MW < 18.1) | Unlinked NDSCR records | Unlinked HES cases |
|  (NDSCR records) | 4939 | 3694 | 449 | 662 | 534 | 137 | ─ |
|  (HES records) | 4941 | 3703 | 446 | 646 | 654 | ─ | 2280 |
| NDSCR record1 |  |  |  |  |  |  |  |
| *East Midlands* | 10.9% | 9.0% | 4.5% | 4.5% | 5.6% | 10.0% | ─ |
| *East of England* | 9.0% | 10.0% | 11.6% | 7.0% | 9.4% | 8.3% | ─ |
| *Greater London* | 20.8% | 19.9% | 27.2% | 32.6% | 28.9% | 45.0% | ─ |
| *North East* | 9.8% | 9.0% | 6.9% | 10.3% | 4.6% | < 8.0% | ─ |
| *North West* | 14.6% | 16.6% | 17.8% | 16.2% | 19.6% | 10.0% | ─ |
| *South East* | 12.7% | 12.4% | 11.9% | 13.6% | 13.2% | < 8.0% | ─ |
| *South West* | 10.3% | 9.5% | 7.4% | 9.5% | 7.6% | < 8.0% | ─ |
| *West Midlands* | 11.9% | 13.5% | 12.6% | 6.4% | 11.0% | 13.3% | ─ |
| HES record1 |  |  |  |  |  |  |  |
| *East Midlands* | 10.8% | 9.1% | 4.1% | 3.3% | 8.0% | ─ | 7.2% |
| *East of England* | 8.9% | 9.9% | 10.5% | 7.6% | 6.0% | ─ | 8.7% |
| *Greater London* | 20.8% | 20.0% | 26.7% | 27.5% | 35.6% | ─ | 26.6% |
| *North East* | 10.0% | 8.9% | 6.8% | 13.3% | 5.3% | ─ | 7.2% |
| *North West* | 14.8% | 16.7% | 18.3% | 25.4% | 16.3% | ─ | 19.6% |
| *South East* | 12.6% | 12.4% | 12.3% | 9.5% | 12.3% | ─ | 13.7% |
| *South West* | 10.3% | 9.5% | 8.9% | 8.3% | 6.0% | ─ | 7.2% |
| *West Midlands* | 11.7% | 13.5% | 12.3% | 5.1% | 10.5% | ─ | 9.8% |
| DOB: Date of birth; HES: Hospital Episode Statistics for England; MW: match weight; NDSCR: National Down Syndrome Cytogenetic Register.NDSCR records exclude those with missing birth outcome. All data are column proportions, ignoring missing data, so that associations between region and linkage quality are reflected by differences in proportion across rows. Probabilistic links are grouped by 'match weight', a score reflecting the level of agreement over matching variables (see Methods). 1The number of candidate links may be higher than the number of records in either file, indicating ambiguity of multiple links with equal agreement; for two of such candidate links, either at least one is false or both are true and it is the records in the contributing files that have not been completely deduplicated.Source: Hospital Episode Statistics (HES), NHS Digital (Copyright © 2019. Re-used with the permission of NHS Digital. All rights reserved) and the National Down Syndrome Cytogenetic Register (NDSCR), Public Health England. |

Figure S10 Proportion of HES cases, by number of episodes in first year of life, and year of birth

Source: Hospital Episode Statistics (HES), NHS Digital (Copyright © 2019. Re-used with the permission of NHS Digital. All rights reserved).

# References

1. Office for National Statistics. *Number of live births at home and total live births, England, 1994 to 2014 birth registrations*; 2016.