International Journal of Epidemiology, 2024, **53(3)**, dyae065
https://doi.org/10.1093/ije/dyae065
Data Resource Profile

Data Resource Profile

Data Resource Profile: A national linked mother-baby cohort of health, education and social care data in England (ECHILD-MB)

Qi Feng [1], Georgina Ireland[1], Ruth Gilbert[1] and Katie Harron [1,*]

[1]Great Ormond Street Institute of Child Health, University College London, London, UK

*Corresponding author. Great Ormond Street Institute of Child Health, University College London, London WC1N 1EH, UK. E-mail: k.harron@ucl.ac.uk

Keywords: Mother-baby cohort, Hospital Episode Statistics, National Pupil Database, data linkage, intergenerational effects.

Key Features

- The Education and Child Health Insights from Linked Data Mother-Baby (ECHILD-MB) cohort was created by linking data from National Health Service (NHS) Hospital Episode Statistics (HES) with the National Pupil Database (NPD), to examine intergenerational effects of maternal exposures on child outcomes in education, health and social care in England.
- We extracted 14.5 million baby records from HES for births between 1 April 1997 and 31 January 2022 and, using a validated mother-baby data linkage algorithm, linked 13.6 million (94.1%) of these to the delivery records of 8.0 million mothers.
- The linked cohort captures 87.3% of all births and 87.7% of all live births in England, and as of 2023, includes mothers aged 12–37 and their children aged 0–24 years. The cohort is representative of national birth statistics.
- All individuals in the cohort are linked to HES and the NPD, which include routinely collected data on sociodemographics, education, health and use of children's social care.
- The ECHILD-MB cohort data is accessible to accredited researchers as part of the ECHILD project [echild.org.uk].

Data resource basics

Background: why is it important?

Maternal physical, psychological and social risk factors extend beyond affecting mothers' individual wellbeing to significantly influence their children. Research consistently shows that maternal exposure to poor nutrition and psychological and social stressors prior to and during pregnancy is associated with an increased risk of adverse birth outcomes, developmental disorders and chronic health conditions of children in both childhood and later life.[1–10] These maternal exposures may even date back to the mother's childhood.[8] Recognizing the intricate interplay of these factors within families and across generations is pivotal for designing targeted interventions to enhance the health and wellbeing of current and future generations.

Longitudinal mother-baby cohort studies are useful in addressing research questions about the influence of maternal exposures on outcomes in their child. There are generally two types of mother-baby cohorts: recruited and consented cohorts, and administrative data-derived cohorts. A key advantage of recruited and consented cohorts is the capture of assessment data from mothers and their children throughout various stages of pregnancy and beyond. In the past decades, many such cohorts have been established.[11] Examples in England include Avon Longitudinal Study of Parents and Children study,[12] 1970 British Cohort Study,[13] Born in Bradford cohort,[14] Southampton Women's Survey[15] and the Gateshead Millennium Study.[16] Some of these cohorts are linked to administrative data for follow up.[12,17] These cohorts collect data on maternal exposures during pregnancy, but often have limited data on maternal exposures prior to pregnancy, such as the mother's childhood, adolescence, and early adulthood. Furthermore, these cohorts can be constrained by relatively small sample sizes, limited follow-up periods or geographical restrictions.[11]

By contrast, administrative data-derived cohorts have distinct advantages, including comprehensive coverage, being less prone to the selection biases, the ability to examine cohorts from past decades and, very recently, the power to examine rare conditions. Several countries have derived national longitudinal administrative cohorts.[18,19]

We derived a national mother-baby cohort nested within the Education and Child Health Insights from Linked Data (ECHILD) Database.[20,21] ECHILD is a linked collection of the National Health Services (NHS) Hospital Episode Statistics (HES) and the Department for Education (DfE) National Pupil Database (NPD) for a population-based cohort of children and young people born in England. In this

Received: 18 December 2023. **Editorial Decision:** 25 March 2024. **Accepted:** 13 April 2024

data resource profile, we describe how we derived the mother-baby cohort, and we present the results of the linkage and its validation, as well as the basic characteristics of the cohort.

Recruitment: mother-baby linkage

HES captures separate records of births for babies and deliveries for mothers in NHS hospitals. However, routine identification of mother-baby pairs is lacking. This section outlines the statistical process used to perform mother-baby linkage using de-identified data, aiming to identify and link baby birth records to maternal delivery records. A concise overview of the method is provided here, with detailed information available in the Supplementary Methods (available as Supplementary data at *IJE* online).

We first identified 14 494 782 birth records and 14 611 863 delivery records between 1 April 1997 and 31 January 2022 from HES data using validated codelists. In HES, delivery episodes for mothers and birth episodes for babies include additional fields on delivery procedures and outcomes, which are called the baby/maternity tail. The baby/maternity tail ideally contains the same information in delivery and birth records, but is sometimes incomplete.

Second, we used a previously validated algorithm for linking delivery and birth records.[22,23] We performed deterministic linkage using seven linking variables (including location, delivery and birth characteristics), and probabilistic linkage using 23 linking variables for the unlinked records. Record pairs with implausible dates were not considered; for example, babies discharged prior to the mother's admission, or mothers discharged prior to the baby's admission.

For record pairs surpassing selected match weight cutoff values, we prioritized the highest match weight for each baby and then for each mother. Mother-baby linkage was conducted individually for each financial year (1 April to the next 31 March) due to varying data quality over time in HES. In probabilistic linkage, matches across financial years were permitted by incorporating maternal delivery data from March of the preceding financial year to April of the subsequent financial year.

Participants: who is included?

In total, 13 635 655 out of 14 494 782 birth records (94.1%) were linked to delivery records (Figure 1). The linkage rate varied by year, with the lowest linkage rate in 1997 (89.5%) and 1998 (89.4%), and the highest linkage rate in 2010 (96.2%); the linkage rate for singleton births was higher than for multiple births (94.4% vs 80.7%) (Supplementary Figure S1, available as Supplementary data at *IJE* online).

Linkage rates also varied by hospital. Among the 341 NHS hospital groups (identified by three-digit hospital codes), 43 groups (birth $n = 9148$) had a 0% linkage rate, 13 groups (birth $n = 55\,846$) 0.1–10.0%, nine groups (birth $n = 14\,024$) 10.1– 50.0% and 93 groups (birth $n = 4\,062\,049$) 50.1–94.1%. The remaining 183 groups (birth $n = 10\,353\,715$, 71.4% of all births) had linkage rates of >94.1%; 22 groups (birth $n = 135\,349$) had linkage rates >99.0%.

Compared with linked babies, the 5.9% of unlinked babies were more likely to have Black or Mixed ethnic background (12.5% vs 10%), to have an older mother (31.0 vs 29.2 years), to be born before or at 37 weeks (24.4% vs 14.3%) and to have lower birthweight (3120 vs 3340 g) (Table 1). Unlinked babies were also more likely to be admitted to special care (17.0% vs 10.3%) or intensive care (6.8% vs 2.7%), or be recorded as being a stillbirth (1.4% vs 0.2%). Similar patterns were observed in singleton births and multiple births (Supplementary Table S1, available as Supplementary data at *IJE* online).

We used the 2021 Community Services Data Set (CSDS) [24] as a gold standard to evaluate the linkage results. Despite representing only a subset of all births, CSDS facilitates valuable linkage validation, as the mother's unique ID is captured on the child's record, which is the same as the IDs used in HES. CSDS comprised records for 2 581 393 babies from 313 local authorities in England. We estimated missed match rates and false match rates.[23,25] Overall, 2 541 772 (98.5%) were linked in this study, and the remaining 39 621 were missed (1.5%). Of the linked records, 2 523 476 were true matches, giving a positive predictive value of 99.3% (2 523 476 true matches/2 541 772 linked records), a false match rate of 0.72% (18 296 false matches/2 541 772 linked records) and a sensitivity of 97.8% (2 523 476 true matches/2 581 393 total mother-baby dyads in CSDS; Supplementary Table S2, available as Supplementary data at *IJE* online).

We evaluated the coverage of mother-baby linkage among all births and live births, relative to the Office for National Statistics (ONS) statistics (1998–2021). The identified birth records covered 92.7% of all births and 92.9% of live births, and the linked babies covered 87.3% of all births and 87.6% of live births (Supplementary Table S3, available as Supplementary data at *IJE* online). We compared the distributions of gestational age, birthweight and maternal age with national birth statistics published by ONS, and observed high levels of agreement (Figure 2, Supplementary Figure S2, available as Supplementary data at *IJE* online).

Data collected

The mother-baby cohort, nested within the ECHILD database,[20,21] shares data sources (NPD and HES) with ECHILD and is de-identified. Both NPD and HES regularly collect and compile information, undergoing quality assurance checks upon submission to DfE and NHS England. NHS England performed the linkage between HES and NPD, with details outlined elsewhere.[26] Briefly, DfE securely transferred identifiers from NPD to NHS England, where deterministic linkage was used to create an anonymized linkage spine connecting the NPD unique identifier (aPMR) to the HES unique identifier (TokenID).

HES contains records for all hospital activity provided or paid for by NHS, including births, inpatient admissions, outpatient appointments, accident and emergency attendances, mortality, demographics and standardized codes for diagnosis and procedures.[27] Data on socioeconomic status are collected at area level using Index of Multiple Deprivation.

NPD contains records related to state-funded education and the use of social care services. NPD data are collected by local authority, but the specification of data that are collected is centralized and determined by the DfE. It includes information from different educational settings about pupils' characteristics, including age, gender, ethnicity, special educational needs and free school meals. Educational outcomes are also documented, including national assessments and examinations, absences, exclusions and participation in post-16 education.[26,28] NPD's social care modules, Children in Need and

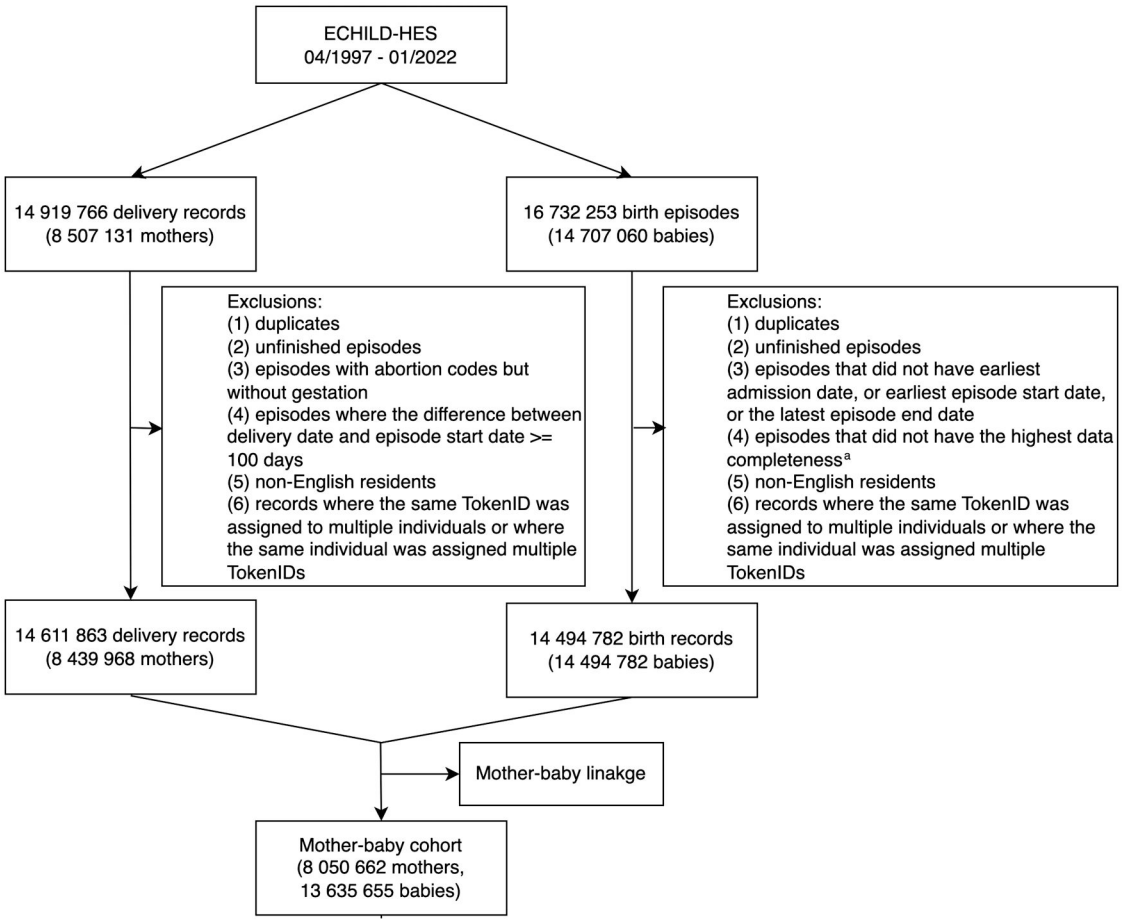

Figure 1. Flow diagram of participant recruitment in ECHILD-MB cohort. For records where the same TokenID (unique individual identifier in HES data) was assigned to multiple individuals, records were considered as different individuals when they had discrepant values in birthweight, maternal age or gestational age. For the records where the same individual was assigned multiple TokenIDs, records were considered as the same individual when they had consistent values in all other variables. [a]In HES, one birth event may take multiple episodes, in which case we included the episode that had the highest data completeness (lowest missing level). HES, Hospital Episode Statistics; ECHILD-HES, Education and Child Health Insights from Linked Data—Hospital Episode Statistics; ECHILD-MB: national linked mother-baby cohort of health, education and social care data in England

Children Looked-after Return, collect information on children using social care services and those in care (looked-after children in the UK) (Table 2). Data on socioeconomic status are collected at area level using the Index of Multiple Deprivation and at individual level using eligibility for free school meals as a surrogate measurement.

Due to variations in age coverage and collection periods across NPD datasets, the availability of specific datasets for mothers and babies is contingent on their respective birth years. Figure 3 demonstrates the availability for a mother born in 1988 and her child born in 2012. To facilitate visualization of data availability for any mother-baby pair, we created an online tool [https://ermadake.shinyapps.io/echild_mb_data/].

We imputed missing values in birth records using the linked delivery record of the mother. We removed implausible values of gestational age and birthweight that fell more than three standard deviations from the average.[29] In case of multiple births, we also imputed missing values in gestational age and maternal age by copying data across multiples. After imputing missing values in baby records using values in mother records, data completeness increased (Supplementary Table S4, available as Supplementary data at *IJE* online).

Data resource uses

The ECHILD-MB cohort serves as a valuable resource for investigating intergenerational effects of maternal exposure before and during pregnancy on children's health, education and social care outcomes. For example, previous research has examined the associations between pre-pregnancy psychosocial risk factors and infant outcomes, finding that exposure to psychosocial risk factors, including teenage motherhood, previous teenage motherhood, a history of adversity, mental health or behavioural conditions, or living in the most deprived quintile during the 2 years before pregnancy, significantly increased children's risk of low birthweight, injury admission and mortality during infancy.[7] The ECHILD-MB will allow this area of research to be further developed due to the inclusion of information on educational and social care contacts.

As the ECHILD-MB cohort matures, the extended follow-up will enable further comprehensive analyses. The cohort extends its utility beyond examining health outcomes at birth and infancy by facilitating investigations into long-term health trends during childhood, adolescence, early adulthood and later life stages. Additionally, the cohort uniquely allows for the exploration of educational outcomes and use of social

Table 1. Characteristics of the linked and unlinked baby birth records

Characteristic	Unlinked babies ($n = 859\,127$)	Linked babies ($n = 13\,635\,655$)	Total ($n = 14\,494\,782$)
Male sex	447 338 (52.1%)	6 968 829 (51.1%)	7 416 167 (51.2%)
Ethnicity			
Asian	58 996 (10.1%)	1 180 172 (11.5%)	1 239 168 (11.4%)
Black	40 051 (6.8%)	554 500 (5.4%)	594 551 (5.5%)
Mixed	33 363 (5.7%)	475 789 (4.6%)	509 152 (4.7%)
White	453 988 (77.4%)	8 031 229 (78.4%)	8 485 217 (78.4%)
Unknown	272 729 (31.7%)	3 393 965 (24.9%)	3 666 694 (25.3%)
Index of multiple deprivation category			
Quintile 1 (most deprived)	26 919 (20.8%)	3 703 342 (27.3%)	3 730 261 (27.2%)
Quintile 2	23 108 (17.9%)	2 908 159 (21.4%)	2 931 267 (21.4%)
Quintile 3	18 219 (14.1%)	2 445 729 (18.0%)	2 463 948 (18%)
Quintile 4	14 787 (11.4%)	2 181 256 (16.1%)	2 196 043 (16%)
Quintile 5 (least deprived)	46 320 (35.8%)	2 342 944 (17.3%)	2 389 264 (17.4%)
Unknown	729 774 (84.9%)	54 225 (0.4%)	783 999 (5.4%)
Maternal age, years, mean (SD)	31.0 (6.38)	29.2 (5.86)	29.3 (5.88)
<20	10 414 (3.8%)	617 308 (5.2%)	627 722 (5.2%)
20–24	34 874 (12.7%)	2 041 857 (17.3%)	2 076 731 (17.2%)
25–29	63 493 (23.1%)	3 264 787 (27.6%)	3 328 280 (27.5%)
30–34	84 327 (30.6%)	3 553 169 (30.1%)	3 637 496 (30.1%)
35–39	59 379 (21.6%)	1 910 529 (16.2%)	1 969 908 (16.3%)
40–44	18 337 (6.7%)	399 069 (3.4%)	417 406 (3.5%)
≥45	4446 (1.6%)	21 360 (0.2%)	25 806 (0.2%)
Unknown	583 857 (68.0%)	1 827 576 (13.4%)	2 411 433 (16.6%)
Gestational age at birth, weeks, median (IQR)	39.0 (38.0–40.0)	39.0 (38.0–40.0)	39.0 (38.0–40.0)
≤27	8977 (3.1%)	29 637 (0.3%)	38 614 (0.4%)
28–32	8112 (2.8%)	104 886 (1.0%)	112 998 (1.1%)
33–37	53 887 (18.5%)	1 309 188 (13.0%)	1 363 075 (13.2%)
38–41	211 454 (72.4%)	8 259 080 (82.2%)	8 470 534 (81.9%)
≥42	9603 (3.3%)	343 465 (3.4%)	353 068 (3.4%)
Unknown	567 094 (66.0%)	3 589 399 (26.3%)	4 156 493 (28.7%)
Birthweight, g, mean (SD)	3120 (798)	3340 (588)	3330 (596)
<1500	15 668 (4.9%)	105 459 (1%)	121 127 (1.1%)
1500–1999	10 290 (3.2%)	146 669 (1.4%)	156 959 (1.4%)
2000–2499	25 017 (7.8%)	529 052 (4.9%)	554 069 (5%)
2500–2999	60 201 (18.8%)	1 969 485 (18.3%)	2 029 686 (18.4%)
3000–3499	103 391 (32.3%)	4 242 935 (39.5%)	4 346 326 (39.3%)
3500–3999	78 027 (24.3%)	3 344 720 (31.1%)	3 422 747 (30.9%)
4000–4499	23 805 (7.4%)	360 763 (3.4%)	384 568 (3.5%)
4500–4999	3534 (1.1%)	34 346 (0.3%)	37 880 (0.3%)
≥5000	614 (0.2%)	5091 (0%)	5705 (0.1%)
Unknown	538 580 (62.7%)	2 897 135 (21.2%)	3 435 715 (23.7%)
Neonatal care admission			
Normal care	368 233 (76.3%)	8 042 021 (87.1%)	8 410 254 (86.5%)
Special care	81 894 (17.0%)	948 222 (10.3%)	1 030 116 (10.6%)
L1 intensive care	20 697 (4.3%)	155 114 (1.7%)	175 811 (1.8%)
L2 intensive care	12 009 (2.5%)	92 457 (1.0%)	104 466 (1.1%)
Unknown	376 294 (43.8%)	4 397 841 (32.3%)	4 774 135 (32.9%)
Stillbirth	11 646 (1.4%)	29 440 (0.2%)	41 086 (0.3%)

When calculating percentage for the Unknown category, the denominator was the total number of babies; when calculating percentage for other categories, the denominator was the total number of babies excluding the unknown category.
IQR, interquartile range; SD, standard deviation.

care services. These findings are instrumental in identifying vulnerable children at heightened risk of adverse outcomes, informing the design of early intervention strategies for targeted and effective support.

Strengths and weaknesses

The ECHILD-MB cohort is characterized with a substantial sample size, seamless integration with national administrative databases, and extensive data encompassing education, health and social care. Further strengths include nationwide coverage and representativeness of mother-baby pairs, good linkage accuracy and the unique capability to explore intergenerational effects of maternal exposures.

The mother-baby cohort encompassed 87.3% of all births and 87.6% of live births in England from 1998 to 2021. The under-ascertainment can be explained by several factors. First, a 5.4% opt-out rate (as of November 2023) among English residents limited data sharing for research or policy purposes.[30] Second, the ONS Birth Registration dataset, mandated to report all births in England, includes records from various settings such as NHS hospitals, private hospitals and homes. Our extraction focused solely on HES birth records in NHS hospitals, resulting in the omission of births not recorded

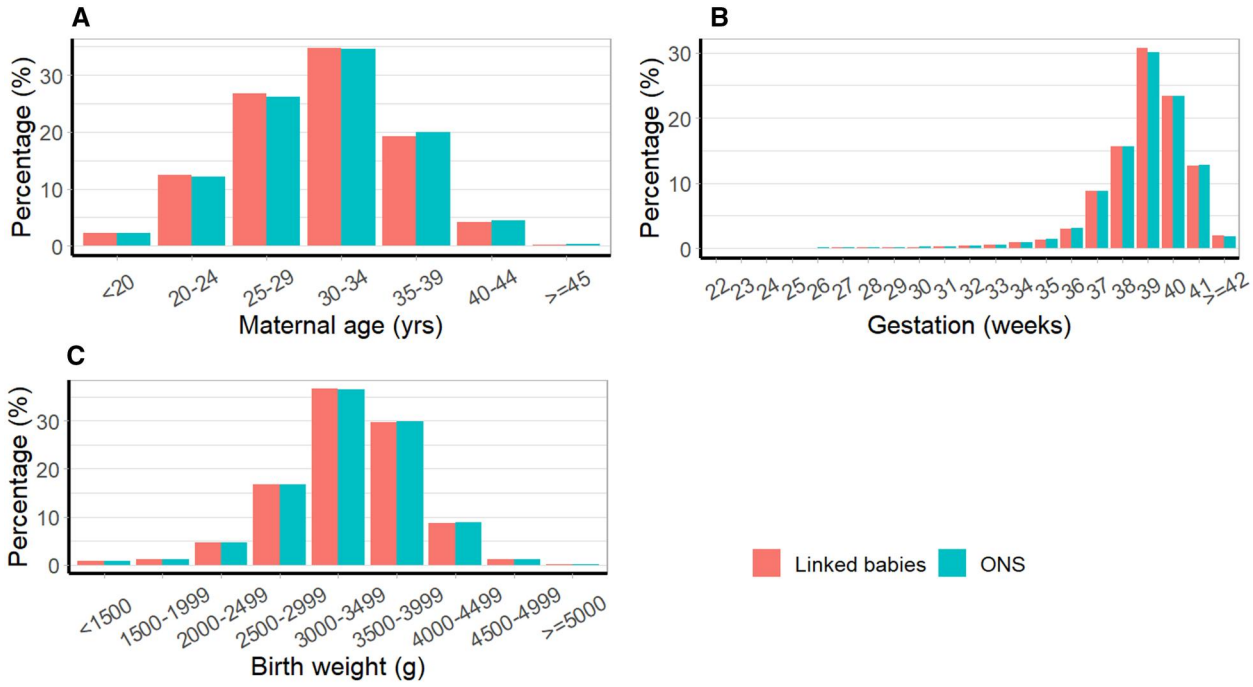

Figure 2. Comparison of the mother-baby cohort with data from the Office of National Statistics (ONS) on maternal age, gestational age and birthweight in the 2021 financial year. Red: ECHILD-MB (national linked mother-baby cohort of health, education and social care data in England) data. Green: Office for National Statistics. Each financial year starts on 1 April and ends on 31 March next calendar year

in HES. Previous work has shown that approximately 95% of Birth Registration records can be linked to HES.[31]

The overall linkage rate in our mother-baby linkage was 94.1%, aligning with a prior study using a smaller HES dataset for linking singleton births.[22] Linkage rates varied across years, hospitals and clinical scenarios, revealing implications for NHS data collection. The linkage rate was correlated with data quality, showing higher rates in later years, consistent with the improving data completeness in HES birth records over time.[32] Hospital-specific variation existed, with 28% of births in hospitals having below-average linkage rates (average linkage rate 87.8%) and lower data completeness/accuracy. Analysis of linked and unlinked babies identified risk factors for non-linkage, including Black or Mixed ethnicities, deprivation and older maternal age, all associated with adverse birth outcomes,[33,34] as observed in this study. Vulnerable individuals and adverse clinical situations yielded poorer data quality, resulting in a lower linkage rate.[35] Whereas valid linkage aids in imputing missing values, enhancing data completeness and accuracy during collection is crucial. The NHS should focus on improving data collection practices and quality at the hospital level for more strategic and efficient outcomes.

Mother-baby linkage accuracy was validated by cross-referencing link status with the CSDS dataset and comparing linked baby characteristics with national statistics. Our accuracy estimates are consistent with a prior study that used a different external dataset (Maternity Information System) for mother-baby linkage validation in 2012.[23] Although the CSDS dataset covered only a subset of identified babies with community service contact, it spanned almost all local authorities in England, ensuring comprehensive coverage. Additionally, linked babies exhibited birth statistics comparable to ONS data, indicating robust national representativeness. Despite these strengths, further validation using external datasets containing actual mother-baby dyads would enhance the reliability of the linkage process.

The ECHILD-MB cohort has inherent limitations stemming from the non-research nature of administrative datasets. This restricts the depth of research possibilities and the interpretation of findings, due to the absence of detailed information on socioeconomic status, genetic background and biomarkers, and potential under-reporting of chronic health conditions. Second, ECHILD-MB cohort solely includes biological mother-baby dyads from delivery and birth records, omitting data on adoption and conception methods. Third, linkage errors accounting for missed links (5.9%) and false links (0.7%), although low, may introduce bias in epidemiological research. The unlinked birth records were more likely to belong to Black or Mixed ethnicities, and those with older maternal age, resulting in under-representation of these characteristics in the linkage cohort. This selection bias is significant, as these characteristics are often associated with the exposure or outcome of interest. Although the false link rate is low (0.7%), it poses the risk of information bias, specifically misclassification bias. Depending on whether false links are equal between exposed and control groups, misclassification bias can be either differential or non-differential, potentially introducing noise and diluting associations in epidemiological analyses. Fourth, this cohort has a major focus on maternal factors and child health, but data on fathers and other family members are limited. Although these data could be an important complement to further examine child outcomes, collecting such data is difficult, particularly using administrative data.[36–38]

Data resource access

ECHILD-MB data is available to external researchers and accessible via ONS Secure Research Service, as part of ECHILD

Table 2. Key data sources included in a national linked mother-baby cohort of health, education and social care data in England (ECHILD-MB)

Data source	Description	Year coverage	Age coverage (years)	Key variables
HES				
Admitted patient care	Diagnoses, operations, operation dates, consultant specialty	1997–22	All	Diagnoses, operations, operation dates etc.
Critical care	Critical care start and end dates, number of days of support by organ group, discharge destination	2006–22	Adults	Diagnosis codes etc.
Accident and emergency/ Emergency Care Services Dataset	Type of attendance, mode of arrival, treatments, duration	2006–22	All	Diagnosis codes etc.
Outpatient	Type of appointment, outcome of appointment, medical staff type seeing patient, duration of elective wait	2002–2022	All	Diagnosis codes etc.
ONS linked mortality death registration	Month and year of death, underlying cause of death	1997–2022	All	Date of death, cause of death etc.
Birth notification	Birth notification	2001–22	Birth	Birth weight, gestational age
Birth registration	Birth registration	1996–22	Birth	Sex, multiple indicator, parents' country of birth and occupation
NPD				
Early Years Census	All 2- to 4-year-olds in state-funded early years care and education	2007–22	2–4	Age, sex, ethnicity, SEN
School census pupil level	All pupils in state-maintained educational settings, excluding hospital schools	2005–22	2–16	Age, sex, ethnicity, SEN, FSM eligibility, language
Pupil referral unit census	All pupils in a PRU (non-mainstream school maintained by the state)	2009–13	2–16	Age, sex, ethnicity, SEN, FSM eligibility, language
Alternative provision census	All pupils in non-mainstream, non-maintained educational settings for whom the state is covering tuition costs	2007–22	2–16	Age, sex, ethnicity, SEN, FSM eligibility
Absences	All pupils in state-maintained educational settings, excluding boarding pupils	2005–22	4–16	Number of absences, numbers that were authorized and unauthorized
Exclusions	All pupils in state-maintained educational settings	2001–21	2–16	Number of fixed period exclusions, number of permanent exclusions
Early Years Foundation Stage profile	All children at the end of the Early Years Foundation Stage of education	2002–19	3–5	Early years practitioner assessment scores
KS1 assessment	All children at the end of KS1	1997–22	5–7	Teacher assessment scores
KS2 assessment	All children at the end of KS2	1995–22	7–11	Teacher assessment scores
KS3 assessment	All children at the end of KS3	1998–13	11–14	Teacher assessment scores
KS4 qualification	All pupils in KS4, including those in private schools	2001–21	14–16	Entry for and attainment in GCSE and equivalent qualifications
KS5 qualification	All pupils in KS5, including those in private schools	2002–21	16–18	Entry for and attainment in A-level and equivalent qualifications
National Client Caseload Information System	All young people aged 16–25 years who have SEN or disability	2010–22	16–25	Post-16 activity; not in education, employment or training indicator
Children in Need Census	Referrals to children's social care and all children in need	2008–22	2–16	Referral date, category of need, start date of child protection plan
Children Looked After Return	All children who are looked after	1991–21	2–16	Placement start and end dates, type of placement setting, legal basis for placement

Information on diagnoses, treatments, and procedures for each episode of care is recorded by clinical coders based on patient care records and/or discharge summaries using standardized codes. In the Admitted Patient Care, Critical Care and Outpatient modules, diagnoses are recorded using the International Classification of Diseases (ICD) version 10, and treatments and procedures are recorded using the Office of Population Censuses and Surveys (OPCS) version 4. In the Accident and Emergency module, bespoke codes are used to record diagnoses and treatments[14]; however, these are much more limited than ICD-10 and OPCS-4 codes. The School Census Pupil Level module is collected on a termly basis in October (Autumn census), January (Spring census), and May (Summer census). The other education census modules are collected in January only.
HES, Hospital Episode Statistics; ONS, Office for National Statistics; NPD, National Pupil Database; PRU, pupil referral unit; SEN, special educational needs; FSM, free school meals; GCSE, General Certificate of Secondary Education (national examinations taken by students at the end of compulsory education); KS, Key Stage.

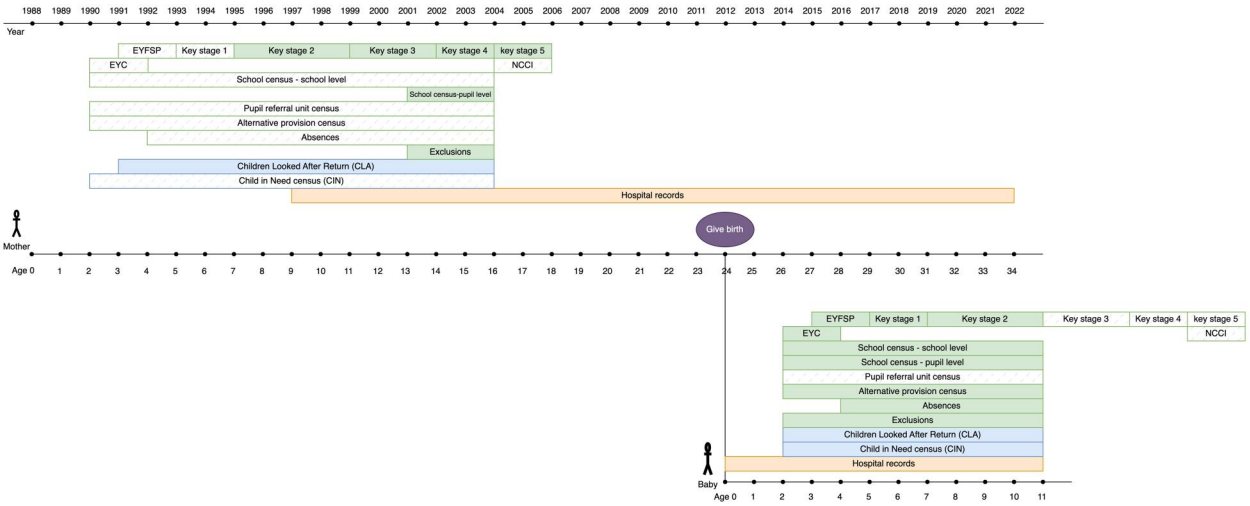

Figure 3. Example of data availability of different National Pupil Database (NPD) datasets for a mother born in 1988 and her child born in 2012. Shaded: data available. Empty: data not available. Green: education datasets. Blue: social care datasets. Orange: health data. EYFSP, Early Year Foundation Stage Profile; EYC, early year census; NCCI, National Client Caseload Information system. We developed an online tool to visualize the data availability for all possible mother-baby pairs [https://ermadake.shinyapps.io/echild_mb_data/]. Data resource profile: a national linked mother-baby cohort of health, education and social care data in England (ECHILD-MB)

database. Interested researchers can apply to the ECHILD Data Access Committee. For more information on ECHILD and application process, please contact [ich.echild@ucl.ac.uk].

Ethics approval

Ethical approval for the ECHILD project was granted by the National Research Ethics Service (17/LO/1494), NHS Health Research Authority Research Ethics Committee (20/EE/0180 and 21/SW/0159) and is overseen by the UCL Great Ormond Street Institute of Child Health's Joint Research and Development Office (20PE16).

Data availability

See Data Resource Access, above.

Supplementary data

Supplementary data are available at *IJE* online.

Author contributions

K.H. and R.G. conceived the research idea. Q.F. and K.H. performed the mother-baby linkage. All authors interpreted the results. Q.F. wrote the first version of the manuscript. All authors critically reviewed and revised the manuscript.

Funding

This work is supported by ADR UK (Administrative Data Research UK), an Economic and Social Research Council (part of UK Research and Innovation) programme (ES/V000977/1, ES/X000427/1 and ES/X003663/1). R.G. was supported by NIHR Senior Investigator Award and Health Data Research UK (HDRUK2023.0029), an initiative funded by UK Research and Innovation, Department of Health and Social Care (England) and the devolved administrations, and leading medical research charities.

Acknowledgements

We thank NHS England and the Department for Education for providing the access to the Hospital Episode Statistics and the National Pupil Database. We thank NHS England for performing the initial data linkages for ECHILD project. The authors would like to thank the wider ECHILD team including Milagros Ruiz, Ruth Blackburn, Matthew Lilliman, Farzan Ramzan, Tony Stone, Vincent Nguyen and Ania Zylbersztejn for their support with data management, and Linda Wijlaars for contributing to data extraction. The authors would also like to thank HDR UK Social and Environmental Determinants of Health Research Driver Programme.

Conflict of interest

None declared.

References

1. Likhar A, Patil MS. Importance of maternal nutrition in the first 1,000 days of life and its effects on child development: a narrative review. *Cureus* 2022;**14**:e30083.
2. Young MF, Ramakrishnan U. Maternal undernutrition before and during pregnancy and offspring health and development. *Ann Nutr Metab* 2020;**76**:41–53.
3. Zhang P, Wu J, Xun N. Role of maternal nutrition in the health outcomes of mothers and their children: a retrospective analysis. *Med Sci Monit* 2019;**25**:4430–37.
4. Canfield M, Radcliffe P, Marlow S, Boreham M, Gilchrist G. Maternal substance use and child protection: a rapid evidence assessment of factors associated with loss of child care. *Child Abuse Negl* 2017;**70**:11–27.
5. Ahmad K, Kabir E, Keramat SA, Khanam R. Maternal health and health-related behaviours and their associations with child health: evidence from an Australian birth cohort. *PLoS ONE* 2021;**16**:e0257188.
6. Hardie JH, Landale NS. Profiles of risk: maternal health, socioeconomic status, and child health: maternal health, inequality, and child health. *J Marriage Fam* 2013;**75**:651–66.

7. Harron K, Gilbert R, Fagg J, Guttmann A, Meulen JVD. Associations between pre-pregnancy psychosocial risk factors and infant outcomes: a population-based cohort study in England. *Lancet Public Health* 2021;**6**:e97–105.
8. Moog NK, Cummings PD, Jackson KL *et al.* Intergenerational transmission of the effects of maternal exposure to childhood maltreatment in the USA: a retrospective cohort study. *Lancet Public Health* 2023;**8**:e226–37.
9. Koen N, Jones MJ, Nhapi RT *et al.* Maternal psychosocial risk factors and child gestational epigenetic age in a South African birth cohort study. *Transl Psychiatry* 2021;**11**:358.
10. MacGinty R, Lesosky M, Barnett W *et al.* Maternal psychosocial risk factors and lower respiratory tract infection (LRTI) during infancy in a South African birth cohort. *PLoS ONE* 2019;**14**:e0226144.
11. Larsen PS, Kamper-Jørgensen M, Adamson A *et al.* Pregnancy and birth cohort resources in Europe: a large opportunity for aetiological child health research. *Paediatric Perinatal Epid* 2013;**27**:393–414.
12. Fraser A, Macdonald-Wallis C, Tilling K *et al.* Cohort Profile: The Avon longitudinal study of parents and children: ALSPAC mothers cohort. *Int J Epidemiol* 2013;**42**:97–110.
13. Sullivan A, Brown M, Hamer M, Ploubidis GB. Cohort Profile Update: The 1970 British Cohort Study (BCS70). *Int J Epidemiol* 2023;**52**:e179–86.
14. Wright J, Small N, Raynor P *et al.*; on behalf of the Born in Bradford Scientific Collaborators Group. Cohort Profile: The Born in Bradford multi-ethnic family cohort study. *Int J Epidemiol* 2013;**42**:978–91.
15. Inskip HM, Godfrey KM, Robinson SM, Law CM, Barker DJ, Cooper C. Cohort Profile: The Southampton Women's Survey. *Int J Epidemiol* 2006;**35**:42–48.
16. Parkinson KN, Pearce MS, Dale A *et al.* Cohort Profile: The Gateshead Millennium Study. *Int J Epidemiol* 2011;**40**:308–17.
17. Connelly R, Platt L. Cohort Profile: UK Millennium Cohort Study (MCS). *Int J Epidemiol* 2014;**43**:1719–25.
18. Ford JB, Roberts CL, Taylor LK. Characteristics of unmatched maternal and baby records in linked birth records and hospital discharge data. *Paediatric Perinatal Epid* 2006;**20**:329–37.
19. Riordan DV, Morris C, Hattie J, Stark C. Family size and perinatal circumstances, as mental health risk factors in a Scottish birth cohort. *Soc Psychiatry Psychiatr Epidemiol* 2012;**47**:975–83.
20. Libuy N, Harron K, Gilbert R, Caulton R, Cameron E, Blackburn R. Linking education and hospital data in England: linkage process and quality. *Int J Popul Data Sci* 2021;**6**:1671.
21. Mc Grath-Lone L, Libuy N, Harron K *et al.* Data Resource Profile: The Education and Child Health Insights from Linked Data (ECHILD) database. *Int J Epidemiol* 2022;**51**:17–17f.
22. Harron K, Gilbert R, Cromwell D, van der Meulen J, Linking data for mothers and babies in de-identified electronic health data. Gebhardt S, editor. *PLoS One*. 2016;**11**(10):e0164667.
23. 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.
24. Fraser C, Harron K, Barlow J *et al.* Variation in health visiting contacts for children in England: cross-sectional analysis of the 2–2½ year review using administrative data (Community Services Dataset, CSDS). *BMJ Open* 2022;**12**:e053884.
25. Harron K, Wade A, Gilbert R, Muller-Pebody B, Goldstein H. Evaluating bias due to data linkage error in electronic healthcare records. *BMC Med Res Methodol* 2014;**14**:36.
26. ECHILD Group. *ECHILD User Guide (Version 2)*. https://www.ucl.ac.uk/child-health/sites/child_health/files/echild_user_guide_v2.pdf (4 April 2024, date last accessed).
27. 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**:1093–93i.
28. Jay MA, Mc Grath-Lone L, Gilbert R. Data Resource: the National Pupil Database (NPD). *Int J Popul Data Sci* 2019;**4**:1101.
29. Cole TJ, Statnikov Y, Santhakumaran S, Pan H, Modi N; on behalf of the Neonatal Data Analysis Unit and the Preterm Growth Investigator Group Neonatal Data Analysis Unit and the Preterm Growth Investigator Group. Birth weight and longitudinal growth in infants born below 32 weeks' gestation: a UK population study. *Arch Dis Child Fetal Neonatal Ed* 2014;**99**:F34–40.
30. NHS Digital. *National Data Opt-Out Open Data Dashboard*. 2023. https://digital.nhs.uk/dashboards/national-data-opt-out-open-data (4 April 2024, date last accessed).
31. Dattani N, Macfarlane A. Linkage of Maternity Hospital Episode Statistics data to birth registration and notification records for births in England 2005-2014: methods. A population-based birth cohort study. *BMJ Open* 2018;**8**:e017897.
32. Zylbersztejn A, Gilbert R, Hardelid P. Developing a national birth cohort for child health research using a hospital admissions database in England: The impact of changes to data collection practices. *PLoS ONE* 2020;**15**:e0243843.
33. Blumenshine P, Egerter S, Barclay CJ, Cubbin C, Braveman PA. Socioeconomic disparities in adverse birth outcomes. *Am J Prev Med* 2010;**39**:263–72.
34. Lean SC, Derricott H, Jones RL, Heazell AEP. Advanced maternal age and adverse pregnancy outcomes: a systematic review and meta-analysis. *PLoS ONE* 2017;**12**:e0186287.
35. Bohensky MA, Jolley D, Sundararajan V *et al.* Data linkage: a powerful research tool with potential problems. *BMC Health Serv Res* 2010;**10**:346.
36. Sharp GC, Schellhas L, Richardson SS, Lawlor DA. Time to cut the cord: recognizing and addressing the imbalance of DOHaD research towards the study of maternal pregnancy exposures. *J Dev Orig Health Dis* 2019;**10**:509–12.
37. Sharp GC, Lawlor DA, Richardson SS. It's the mother!: How assumptions about the causal primacy of maternal effects influence research on the developmental origins of health and disease. *Soc Sci Med* 2018;**213**:20–27.
38. Lut I, Harron K, Hardelid P, O'Brien M, Woodman J. 'What about the dads?' Linking fathers and children in administrative data: a systematic scoping review. *Big Data Soc* 2022;**9**:205395172110692.

International Journal of Epidemiology, 2024, 53, 1–8
https://doi.org/10.1093/ije/dyae065
Data Resource Profile