ECHILD User Guide

Release: ECHILD User Guide v3.2 - May 2026

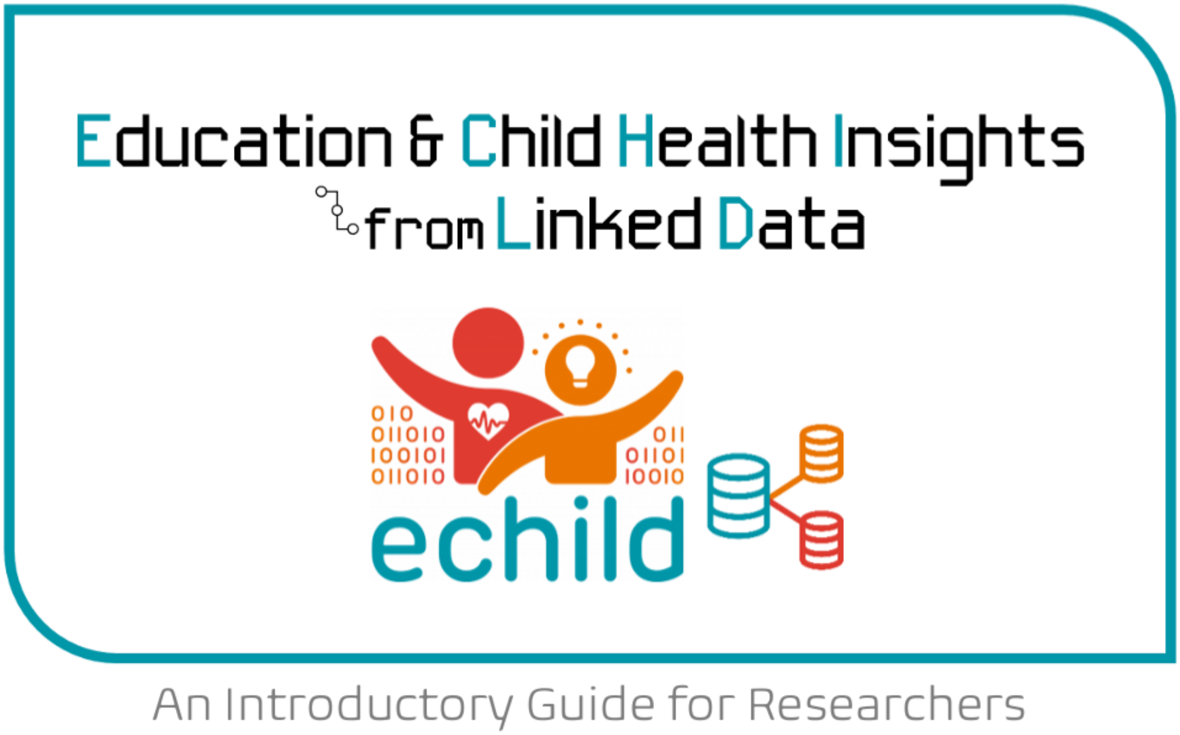
Table of contents

Welcome

Welcome to the ECHILD User Guide. If you have any feedback or spot any errors, you can use the "Report an issue" link present on each page.

The ECHILD research database joins together existing health, education and social care information for all children in England for the first time.

The ECHILD project is led by University College London Great Ormond Street Institute of Child Health in collaboration with the London School of Hygiene & Tropical Medicine and the Institute for Fiscal Studies, in partnership with the Department of Health and Social Care and the Department for Education, working with NHS England and the Office for National Statistics.



Citing this User Guide

When citing this User Guide, you should use the following:

```
Ramzan F, Mc Grath-Lone L, Blackburn R, Gilbert R, Jay M, Lewis K, Lilliman M,
Nguyen V, Ruiz Nishiki M, Stone T, Harron K. Education and Child Health
Insights from Linked Data (ECHILD): an introductory guide for researchers. UCL
(University College London) 2023. doi: 10.5281/zenodo.10854355
```

BibTex reference:

```
@manual{echild2023, author = {Ramzan, F. and McGrath-Lone, L. and Blackburn,
R. and Gilbert, R. and Jay, M. and Lewis, K. and Lilliman, M. and Nguyen, V.
and Ruiz Nishiki, M. and Stone, T. and Harron, K.}, title = {Education and
Child Health Insights from Linked Data (ECHILD): an introductory guide for
researchers}, date = {2023}, doi = {10.5281/zenodo.10854355}, url =
{https://docs.echild.ac.uk/}, publisher = {UCL (University College London)} }
```

ℹ Note

The DOI provided in the above snippets will resolve to the most recent release of this ECHILD User Guide available at the time that the DOI is **accessed**.
If you wish to link to the current release of the ECHILD User Guide at the current time, you should use the DOI shown here: DOI: 10.5281/zenodo.10854355
To find the DOI of a previous release, visit the ECHILD User Guide's Zenodo record and select the applicable version.

1 Introduction

This User Guide aims to give researchers and analysts an introduction to **ECHILD**. It should be read in conjunction with the ECHILD Data Catalogue. If you have any feedback or spot any errors, you can use the “Report an issue” link present on each page.

1.1 Education & Child Health Insights from Linked Data (ECHILD)

ECHILD is a collection of linked, longitudinal, administrative datasets from NHS hospitals, state school education, and children’s social care services for the whole population of children and young people in England. The creation of ECHILD was led by University College London in partnership with NHS England (NHSE) and the Department for Education (DfE), and funded by ADR UK (ECHILD, 2023; Administrative Data Research UK, 2023).

The administrative datasets that have been brought together in ECHILD are well-documented by data providers and the research community. This user guide does not duplicate existing detailed information about these source datasets. Instead, we highlight key aspects that have implications for the potential uses, strengths and limitations of ECHILD. Variables that are included in ECHILD are available separately in the ECHILD Data Catalogue.

1.2 Approvals

Approvals to create and evaluate ECHILD as reported in this user guide were granted by DfE (DR200604.02) and NHSE (DARS-NIC-381972). 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). Further details, including the associated privacy notice, are on the ECHILD website. ECHILD is available for re-use through the Office for National Statistics Secure Research Service (ONS SRS).

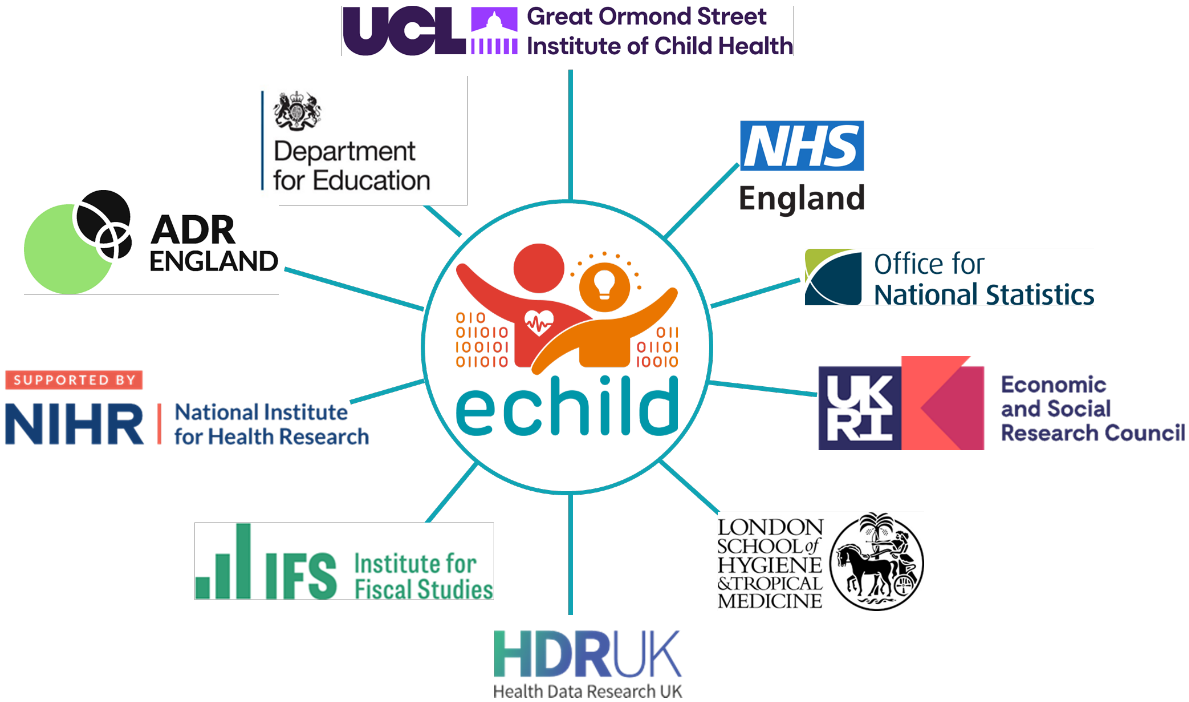


Figure 1: ECHILD partners and contributors

2 Information scope

In England, information on a child's journey through education and social care is recorded in administrative records held by the Department for Education (the National Pupil Database; NPD). NHS England holds information about all NHS hospital contacts (captured in Hospital Episode Statistics; HES). HES records are generated for the purposes of service delivery, e.g., to support financial reimbursement for treatment relating to a hospital stay.

Within ECHILD, healthcare, education and social care records have been linked to create a longitudinal database that follows children over time. The database is very useful for research as health, education and social care trajectories are strongly interrelated from childhood to adulthood. ECHILD provides a valuable opportunity to explore these relationships and to generate evidence for policy and practice (Mc Grath-Lone, Libuy, Harron, *et al.*, 2021).

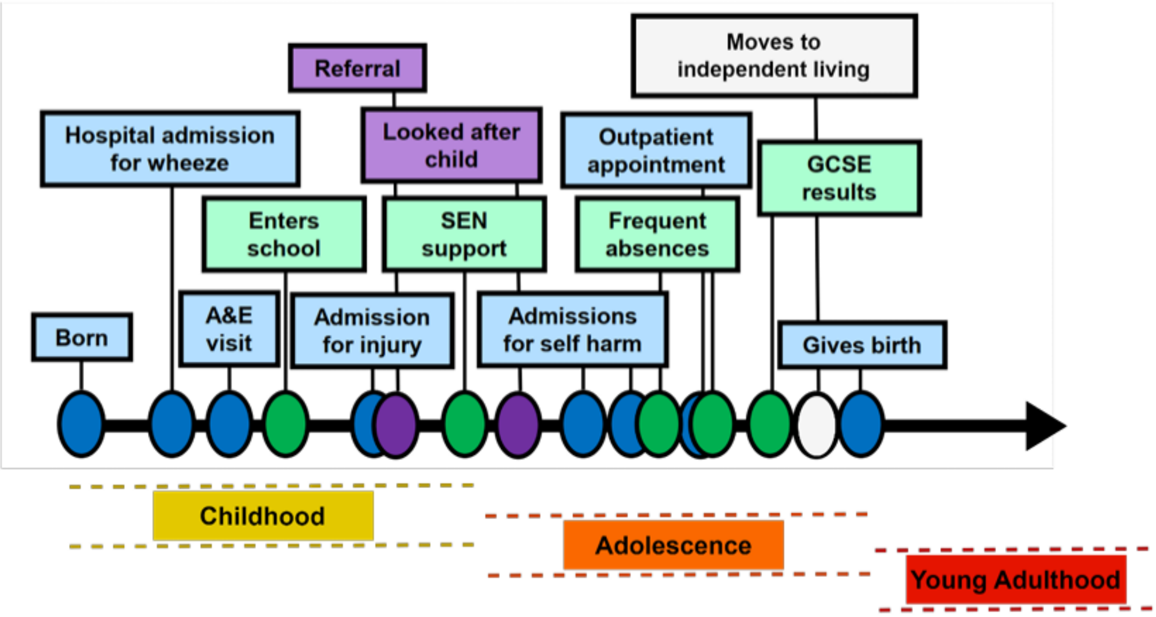


Figure 2: Example trajectory of an individual from birth to young adulthood, captured in the ECHILD data

3 Permitted use

ECHILD can only be used for research that has a clear public benefit to improve the health and well-being of children and young people in England and Wales. Specific research purposes (permitted uses), with examples of relevant research questions, are:

1. **Informing preventative strategies by Healthcare and Education services** e.g., do disabled children attending schools, or living in areas that provide a good level of disability support in school or through social care services, have lower rates of unplanned hospital contacts compared with less supportive schools/areas?
2. **Informing children and their parents** e.g., about variation in special educational needs support and outcomes for children with chronic health conditions or disability.
3. **Informing education and clinical practice** e.g., investigating whether associations between chronic health conditions and lower school attainment are explained by school absence.
4. **Identifying groups who could benefit from intervention** e.g., what are the health outcomes of children post age 16 who have contact with social care services or have special educational needs?
5. **Understanding the most effective methods for working with linked health and education data** e.g., what are the most effective methods for working with linked health and education data?

4 Key Features

- Population-based cohort of children & young people born from September 1984 onwards in England (with planned annual updates)
- Longitudinal data from birth to mid-adulthood (e.g., individuals born in 1984 can be followed up until age 38 in 2022), representing >350 million person-years of follow up
- Pseudonymised datasets that do NOT include any identifiable information (no name, address, postcode, date of birth, Unique Pupil Numbers or NHS numbers).
- Data from health (including information from NHS hospitals, maternity services, mental health services, community services), education (state schools in England) and children's social care, combined with civil birth and death registrations
- Near-population coverage with linked data for an estimated 21 million individuals, with high linkage rates: 99% of pupil records are linked to a health record (National Data Opt Outs are applied to the health data, so effective linkage rates are lower)
- Linkage within and across datasets via a bridging file containing pseudonymised IDs for health and education datasets
- Access to approved researchers in the Office for National Statistics Secure Research Service
- Statistical Disclosure Control performed by the ONS team to ensure that confidentiality is preserved; researchers are not permitted to re-identify individuals

Part I

Data Sources

5 Health data

The data collections detailed below are supplied by NHS England. NHS England is responsible for the collation of all these except for Civil birth and death registrations, which are collated by the Office for National Statistics (ONS).

A pseudonymized, unique identifier (the Token Person ID; TPI) can be used to follow individuals within and across the health datasets.

5.1 Hospital Episode Statistics

Hospital Episode Statistics (HES) captures activity within NHS-funded hospitals in England, including inpatient hospital attendances, outpatient attendances, and A&E contacts. Approximately 97% of all children born in NHS-funded hospitals in England have a birth record in HES.

HES records contain basic demographic information (e.g., sex, ethnicity), area-level deprivation measured by the Index of Multiple Deprivation (IMD) of residential postcode (available at the Medium Super Output Area; MSOA level). Clinical information is based on International Classification of Diseases 10th Revision (ICD-10) diagnostic codes and Office of Population Censuses and Surveys Classification of Interventions and Procedures 4th Revision (OPCS-4) procedure codes (NHS England, 2023g).These codes can be used to create phenotypes; published examples of these are available in the ECHILD phenotype repository.

HES data are collected from hospital care providers for each **financial year** (1st April to 21st March) and curated on an ongoing basis. Following processing and quality assessment, finalised datasets are released for secondary use and remain unchanged thereafter (Boyd *et al.*, 2018).

5.1.1 Admitted Patient Care

Years available: 1997 to 2023

APC is an episode-level dataset of patients admitted for treatment (i.e., requiring the use of a hospital bed), at NHS hospitals in England; includes delivery and birth data, up to 20 diagnostic codes per episode and procedure codes (Herbert *et al.*, 2017; Boyd *et al.*, 2018; Health Data Research, 2023b).

Records within APC are called 'hospital episodes', and each episode relates to a period of care for a patient under a single consultant (consultant episode) within one hospital provider (Herbert *et al.*, 2017; Boyd *et al.*, 2018; Health Data Research, 2023b). The time from initial admission to discharge is called a 'spell', defined as 'periods of continuous care in one provider institution' and each admission spell can be made up of many episodes. A hospital admission commences when a patient is initially admitted for care and ends when a patient is discharged, transferred, or dies. One spell of admission encompasses multiple episodes of care under a different consultant. APC data contain 'Finished Admission Episodes' which are the first episode in a spell of care, and 'Finished Consultant Episodes', which are a continuous period of care under one consultant. Only Finished Consultant Episodes concluded within the financial year (from 1st April to the end of 31st March) are included. Patients with an unfinished consultant episode in the current financial year will have their record represented as a finished episode in the next financial year of HES data.

A hospital admission commences when a patient is initially admitted for care and ends when a patient is discharged, transferred, or dies. One spell of admission encompasses multiple episodes of care under a different consultant.

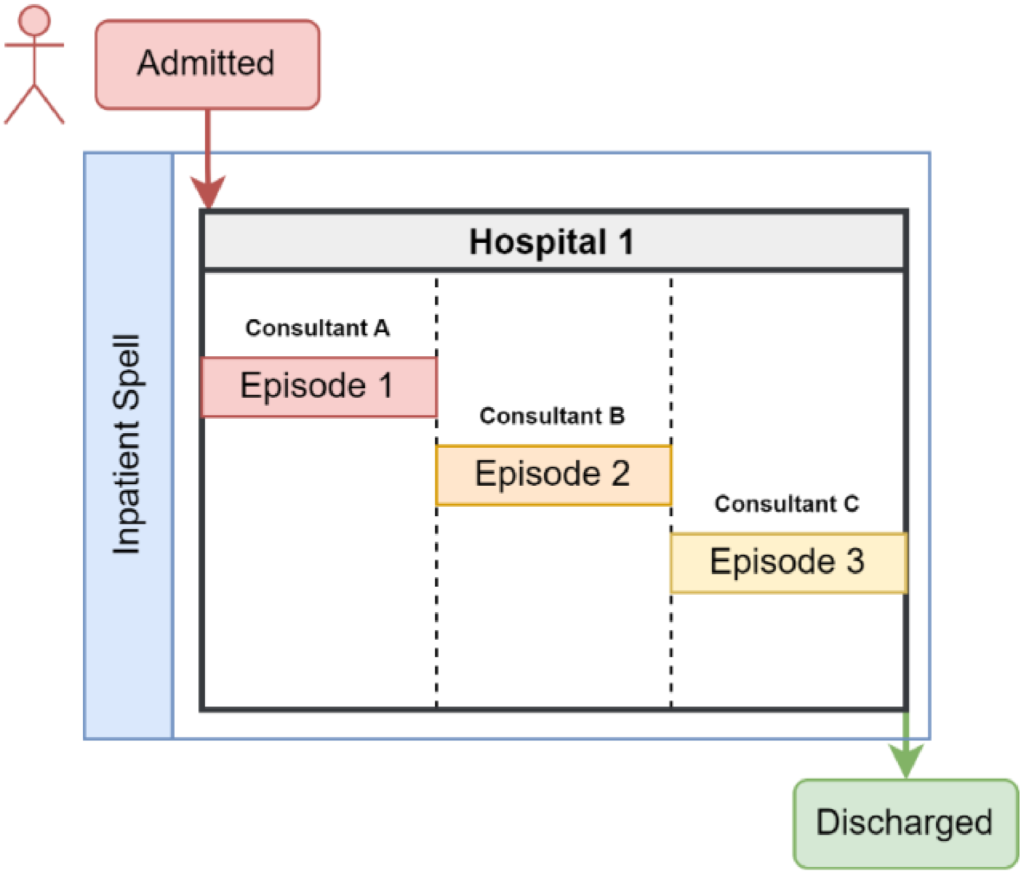


Figure 3: Inpatient spell (single hospital)

In some instances, admitted patients who require specialised treatment may be transferred from one hospital to another more specialist hospital, e.g., transfer to a Children's hospital.

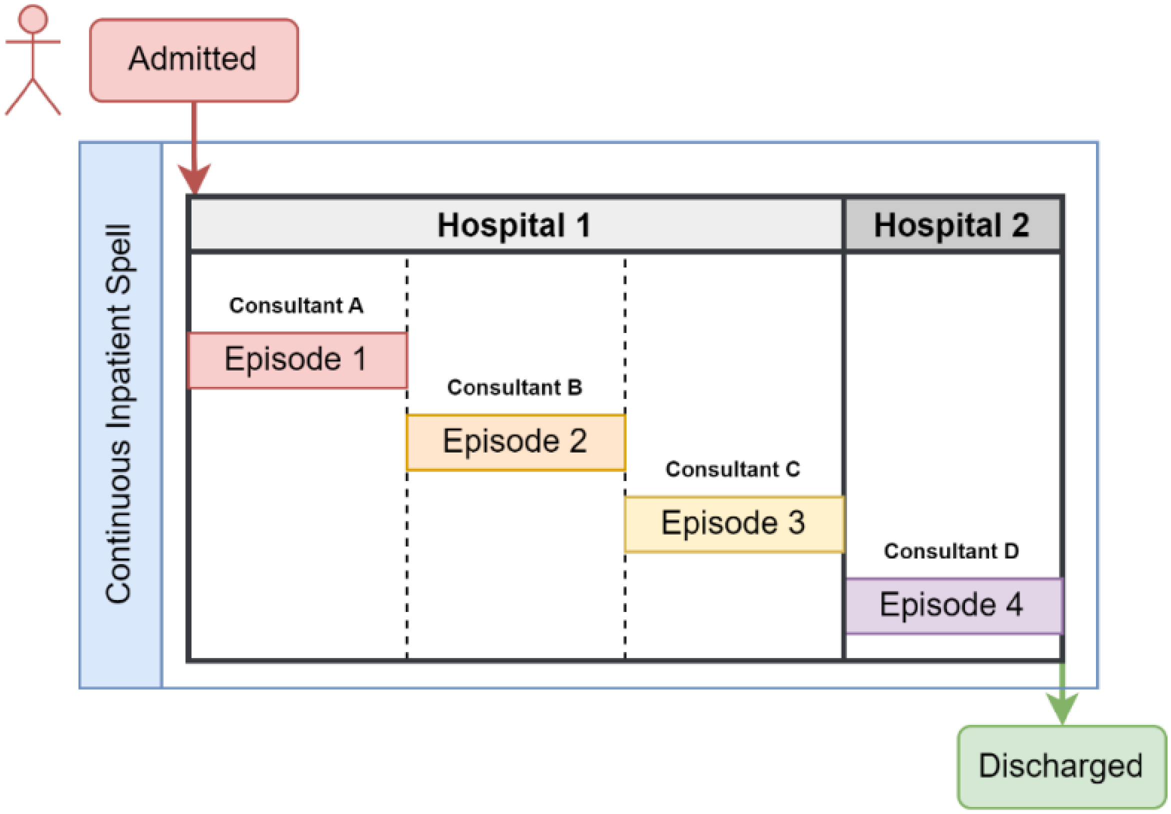


Figure 4: Continuous inpatient spell (across multiple hospitals). In some instances, admitted patients who require specialised treatment may be transferred from one hospital to another more specialist hospital, e.g., transfer to a Children's hospital.

5.1.2 Critical Care (CC)

Years available: 2008 to 2023

Episode-level dataset of patients admitted for treatment and receiving Critical Care (intensive care or high dependency care) at NHS hospitals in England. Data are restricted those treated in wards where constant support and monitoring are required to maintain at least one organ (i.e., an Intensive Care or High Dependency Unit) (Health Data Research, 2023c; NHS England, 2024b). Children are only included in the data from 2017/18 onwards; prior to those, data were restricted to patients on adult designated wards.

5.1.3 Outpatients (OPA)

Years available: 2003 to 2023

Outpatient appointments at English NHS hospitals and NHS commissioned activity in the independent sector (regardless of whether the appointment was attended or not) (NHS England, 2023i; Health Data Research, 2023d). In 2003/04, the OP module was considered experimental and did not have complete coverage as not all providers completed data submissions. Diagnoses and procedures have very low levels of completeness in the Outpatient dataset (5% and 26%, respectively) (Boyd *et al.*, 2018).

5.1.4 Accident & Emergency (A&E) / Emergency Care Dataset (ECDS)

Years available: 2007 to 2020 (A&E) and 2017 2023 (ECDS)

Attendance level dataset collecting information about the treatment received by patients attending A&E Departments, Minor Injury Units and Walk-In Centres, in England (NHS England, 2023e; Health Data Research, 2023a). Bespoke codes are used to record diagnoses and treatments (NHS England, 2023f); however, these are much more limited than ICD-10 and OPCS-4 codes and do not have high levels of completeness. Prior to 2012/13, the A&E module was considered experimental and did not have complete coverage as not all providers completed data submissions (NHS England, 2023e). Bespoke codes are used to record diagnoses and treatments (NHS England, 2023f); however, these are much more limited than ICD-10 and OPCS-4 codes. Prior to 2012/13, the A&E module was considered experimental and did not have complete coverage as not all providers completed data submissions (NHS England, 2023e).

In 2018, the A&E dataset was replaced by the ECDS which is now the national dataset for Urgent & Emergency care (NHS England, 2023d, 2023c; Health Data Research, 2023a). Please note diagnoses aren't consistently captured in ECDS records. ECDS uses a subset of SNOMED CT codes to record diagnoses and treatments (amongst other variables) (NHS England, 2023d, 2023c).

5.1.5 Mental Health Data

5.1.5.1 Mental Health Services Dataset (MHSDS)

Years available: 2016 to 2022

Patient-level dataset that records all activity relating to patients who receive assessments and treatment from Mental Health Services in England, where the patient has (or is thought to have): either a mental health condition; a need for support with their mental well-being; a learning disability; autism; or any other neurodevelopmental condition (NHS England, 2023l).

MHSDS includes data supplied by both Children and Young People's (formerly known as Child and Adolescent) mental health services, and adult mental health services.

MHSDS Data are collected about people treated in NHS-funded services for mental health and wellbeing, learning disability, autism or other neurodevelopmental conditions. Data collected include the type of care received, the length of care episodes, information on contact with mental health and social care professionals, and measures of health and social functioning. Please be reminded that data quality for MHSDS has not yet been fully assured; there is no single, reliable "diagnosis" variable for each individual or care episode. An individual can be associated with referrals to multiple services at the same time, for one or multiple reasons. Having a referral does not necessarily mean the referral is valid, accepted, or that treatment has been received. Further triangulation between MHSDS and other relevant data sources (such as HES, NHSBSA medicine dispensing, SCMD, and third-sector MH service providers) on diagnostic, crisis team, treatment and other service use is needed.

5.1.5.2 Mental Health & Learning Disabilities Dataset (MHLDDS)

Years available: 2014 to 2016

MHLDDS is collected from adult mental health services (NHS England, 2024c) and does not include information on children.

5.1.5.3 Mental Health Minimum Dataset (MHMDS)

Years available: 2006 to 2014

The MHMDS was the preliminary dataset capturing data about the use of adult Mental Health Services in England. However, the MHMDS was superseded by the Mental Health & Learning Disabilities Data Set, which in turn was superseded by the MHSDS (NHS England, 2024c). MHMDS includes data supplied by adult mental health services only and does not include children.

5.1.6 Maternity and Births

5.1.6.1 Maternity Services Data Set (MSDS)

Years available: 2015 to 2019 (v1.5): and 2019 to 2022 (v2.0)

The MSDS is a patient-level dataset that captures information about activity carried out by Maternity Services relating to a mother and baby(ies), from the point of the first booking appointment until the mother and baby(ies) are discharged from maternity services. The MSDS collects records of each stage of the maternity service care pathway in NHS-funded maternity services and includes information not recorded in HES (NHS England, 2023a; Health Data Research, 2023e).

5.1.6.2 Birth notifications

Years available: 2003 to 2022

Birth notifications are documents completed by the doctor or midwife present at birth occurring in an NHS facility. At this point, the baby's NHS Number is issued as part of the 'statutory notification of birth' (but is not visible within ECHILD). Birth notification data includes information that is not found in the birth registration data such as gestation age and ethnicity of the baby as stated by the mother (NHS England, 2023b).

5.1.6.3 Civil Registrations of Births

Years available: 1996 to 2023

Birth Registration Data includes information recorded when births are registered. It is a legal requirement to register all births in England and Wales. These registrations are lodged with a local registry office and collated by the General Register Office (for England and Wales). The General Register Office shares these data with the Office for National Statistics which processes these data and further shares the data with NHS England. All registered live births are included except very late registrations received more than 14 months after the end of each reference year (there are fewer than 100 of these for any given year) (Office for National Statistics, 2024). Birth Registration Data in ECHILD does not contain records relating to stillbirths due to how the ECHILD cohort is defined.

5.1.7 Civil Registrations of Deaths

Years available: 1997 to 2024

It is a legal requirement to register all deaths in England and Wales. Death registrations are lodged with a local registry office and collated by the General Register Office (for England and Wales). The General Register Office shares these data with the Office for National Statistics which processes these data and further shares the data with NHS England. Mortality data contains information taken from death certificates for all deaths registered in England and Wales and includes cause of death, date and place of death. Information related to stillbirths is not available in this data. Some deaths, particularly those referred to a coroner, may not be registered until long after the date of death.

5.1.8 Community Services Dataset (CSDS)

The CSDS captures activity data about children and adults collected by Community Services, including health visiting teams. Such activities may take place in settings such as Health centres; Day care facilities; Schools or Community centres; Mobile facilities, or a patient's own home (NHS England, 2023j). Data are collected about children and adults e.g., personal, demographic or social circumstances; breastfeeding and nutrition; long-term conditions (disabilities), diagnoses and scored assessments. The CSDS is comprised of patient-level data from all publicly funded community services providers e.g., Foundation or Non-Foundation Trusts; Acute Trusts; Mental health Trusts, Community Healthcare Trusts, Independent sector providers and Local Authorities. CSDS data are inconsistently captured over time; meaningful interpretation of data prior to 2018/19 is challenging (Clery *et al.*, 2024).

6 Education and social care data

The data collections detailed below are collated by the Department for Education (DfE).

The National Pupil Database (NPD) was formally established in 2002 based on including a pupil census record for every child in state education (Jay, Mc Grath-Lone and Gilbert, 2019; Department for Education, 2023b). The NPD is made up of modules of data that are collected by the DfE from schools, local authorities and exam-awarding organisations on an ongoing and statutory basis. Information collected as part of NPD is used for funding purposes, policy-making, generating statistics and research. The NPD modules included in ECHILD can be broadly grouped as pupil characteristics, educational outcomes and social care.

NPD does not include information on pupils within private schools, or those being home-schooled, except in relation to public examinations (key stage 4 and 5). Approximately 7% of children (Jay, Mc Grath-Lone and Gilbert, 2019) are enrolled in a private school each year with up to 11% ever enrolled in a private school during their school career (Green *et al.*, 2017).

When a pupil first attends a state-funded school in England e.g., nursery or primary school, or has an education, health and care plan (EHCP) put in place, they are allocated a 'Unique Pupil Number' (UPN), which remains with the pupil throughout their school career regardless of any change in school or local authority (Department for Education, 2019). Social care data is included in the NPD for children who have a UPN. Children receiving social care preschool entry who never have social care during their school years are therefore not included in ECHILD.

UPNs facilitate the transfer of school-based education and attainment data between schools, local authorities and central government and are stored within the NPD. Within ECHILD, a nationally unique and anonymised child-level identifier called the Anonymised Pupil Matching Reference (aPMR) can be used to link data across different years of data collection (Jay, Mc Grath-Lone and Gilbert, 2019).

NPD data are provided for each academic year (1st August to 31st July). Regarding the labelling structure in the following sections, for example, the year 2001/2002 is labelled as 2002.

6.1 Pupil enrollment data collections

6.1.1 School Census Pupil Level

Years available: 2001 to 2022 (Missing: 2020)

The School Census Pupil Level holds information on pupils enrolled in state-funded schools, including local authority-maintained schools, academies, city technical colleges and special schools.

The census modules of the NPD are recorded at the enrolment level. Children who are registered in more than one educational setting will have multiple records in a census. The census modules of the NPD contain information such as age, gender, ethnicity, special educational needs (SEN) support, first language and free school meals (FSM) eligibility.

Data is collected termly: Autumn (October); Spring (January); Summer (May). From Spring 2013/14, the school census also includes pupils enrolled in Pupil Referral Units (previously collected in a separate census).

The census does not contain information for pupils enrolled in hospital schools or non-maintained independent schools (e.g., private schools) or who pursue an apprenticeship, traineeship, training, or work as part of their post-16 options (Department for Education, 2022, 2024b).

After KS4, the School Census Pupil Level module only contains information for young people who continue in full-time post-16 education in schools or colleges.

6.1.2 Early Years Census

Years available: 2008 to 2022

The Early Year Census holds data on children (all 2- to 4-year-olds) in state-funded early years care in any private, voluntary, and independent (PVI) sector nursery, with one or more children receiving funding from DfE. Data collected annually (January) (Department for Education, 2024c).

The census modules of the NPD are recorded at enrolment level. Children who are registered in more than one educational setting will have multiple records in a census. The census modules of the NPD contain information such as age, gender, ethnicity, special educational needs (SEN) support, first language and free school meals (FSM) eligibility.

The Early Years Census only collects information for children who are taking up a government-funded place (Department for Education, 2024c). All 3- and 4-year-olds in England are entitled to government funding; however, funding for 2-year olds is only for specific groups of children, such as those who are in care, who have an education, health and care plan, or whose parents are in receipt of certain benefits. Early years settings that do not have any children who receive direct government funding are not required to submit information via the Early Years Census.

6.1.3 Pupil Referral Unit Census (PRU)

Years available: 2010 to 2013

Information on pupils enrolled in PRUs (a form of school for pupils unable to attend mainstream schools due to factors such as behavioural issues). From Spring 2013/14, these data are collected as part of the School Census.

As with all census modules of the NPD, these are recorded at enrolment level. Children who are registered in more than one educational setting will have multiple records in a census. The census modules of the NPD contain information such as age, gender, ethnicity, special educational needs (SEN) support, first language and free school meals (FSM) eligibility.

6.1.4 Alternative Provision Census (AP)

Years available: 2008 to 2022

The Alternative Provision Census is an annual census (collected in January) of pupils who are educated in alternative provision placements (Department for Education, 2024b). Provision must be arranged by the local authority or school otherwise the child would not receive suitable education e.g., due to illness or if they received a fixed-term exclusion.

The census modules of the NPD are recorded at enrolment level. Children who are registered in more than one educational setting will have multiple records in a census. The census modules of the NPD contain information such as age, gender, ethnicity, special educational needs (SEN) support, first language and free school meals (FSM) eligibility.

6.2 School level data collections

6.2.1 School Level Census

Years available: 2005 to 2022

Collects information from primary schools, secondary schools, special schools, maintained nurseries and academies and pupil referral units three times a year. Private schools are not included.

6.2.2 Get Information about Schools (GIAS)

Years available: 2020 to 2022

GIAS (formerly ‘Edubase’) is the DfE’s public register or dataset containing school characteristics. GIAS is updated whenever a school updates their details. GIAS also maintains information for several organisation types and is used by the DfE to contact establishments, update systems, perform analysis and inform policy decisions (Department for Education, 2024d).

6.3 Educational Outcomes data collections

6.3.1 Early Years Foundation Stage Profile (EYFSP)

Years available: Academic years from 2003 to 2019, and 2022 (Missing: 2020)

The Early Years Foundation Stage Profile contains information on statutory assessment of children in the final year of the Foundation Stage (Reception year).

6.3.2 Key Stage 1 (KS1)

Years available: Academic years from 1998 to 2019, and 2022 (Missing: 2020 to 2021)

Information on assessment of learners by the end of year 2 of schooling (age 7).

6.3.3 Key Stage 2 (KS2) - Exam and Pupil

Years available: Exam – Academic years from 2007 to 2019, and 2022 (Missing: 2020 to 2021); Pupil – Academic years from 1996 to 2019, and 2022 (Missing: 2020 to 2021)

Information on assessment of learners by the end of year 6 of schooling (age 11). 2009/10 Industrial action - teacher boycott affected KS2 SATs (approx. date: May 2010).

6.3.4 Key Stage 3 (KS3) – Pupil and Teacher Assessment

Years available: KS3 – Academic years from 1999 to 2006; Pupil – Academic years from 2007 to 2008; Teacher Assessment – Academic years from 2009 to 2013

Information on assessment of learners by the end of year 9 of schooling (age 14).

6.3.5 Key Stage 4 (KS4) – Exam and Pupil

Years available: Exam – Academic years from 2002 to 2021; Pupil – Academic years from 2002 to 2021

Information on the assessment of learners by the end of year 11 of schooling (age 16). Attainment data are collected for all pupils when they complete nationally recognised assessments at KS4, including those in private schools and further education state sector colleges (Department for Education, 2015).

6.3.6 Key Stage 5 (KS5) – Exam and Student

Years available: Exam – Academic years from 2003 to 2021; Student - Academic years from 2003 to 2021

Information on post-16 assessment of learners in school, sixth forms and Further Education colleges. Attainment data are collected for all pupils when they complete nationally recognised assessments at KS5. Attainment data are collected for all pupils when they complete nationally recognised assessments at KS4 and KS5, including those in private schools and further education state sector colleges (Department for Education, 2015).

6.3.7 Absences

Years available: For Term 2, Academic year 2022 (only); For Term 3, Academic years from 2006 to 2019, and 2022 (Missing: 2020 to 2021)

Information on authorised and unauthorised absences, including reasons for absence, derived from the termly School Census, for 4- to 15-year-olds. Absence data are not recorded for boarding pupils (Department for Education, 2023e). Schools provide information about the reasons for absences (e.g., due to illness, medical appointments, etc) though approximately 1% of schools are able only to provide overall authorised and unauthorised absences (Department for Education, 2023e). Absence data were first collected for 4-year-olds in the 2012/13 academic year. In 2012/13, the period of collection of absence data was also extended to the end of the summer term. In previous years, absence data were only collected for the first half of the summer term. Absence data include ‘persistently absent’ indicators (the threshold for which varies by academic year, though users can specify their own thresholds).

6.3.8 Exclusions

Years available: Academic years from 2002 to 2022

Information on pupil fixed term and permanent exclusions as collected in the termly School Census. Compared to other modules of the NPD, there is a lag in data availability related to exclusions. Data are made available in the summer for the preceding academic year; for example, data for the academic year 2017/18 is released in summer 2019 (Department for Education, 2017).

6.3.9 National Client Caseload Information (NCCIS)

Years available: Academic years from 2013 to 2022

Information on employment destinations. Unlike the School Census Pupil Level module (which only contains information for young people who continue in full-time education post-16), NCCIS includes information about the post-16 activities of all young people aged 16 to 19 years (or aged 16 to 24 years for young people with a current Education, Health and Care (EHC) plan) (Department for Education, 2023d). This activity information is collected by local authorities and used by the DfE to estimate Not in Education, Employment or Training (NEET) rates for young people in England.

6.4 Social Care data collections

The CLA and CIN modules of the NPD contain two different identifiers. The first is an encrypted version of the identifier assigned by the local authority (child ID) that allows social care records for the same individual to be linked over time. However, this identifier is local authority specific and so it is not possible - using this identifier - to link records for the same individual across different local authorities (Mc Grath-Lone, Harron, *et al.*, 2016; Emmott, Jay and Woodman, 2019).

The second is the aPMR, based on Unique Pupil Numbers (UPNs) where available. UPNs have been returned to DfE by local authorities in the CLA module from 1 April 2005 and in the CIN module from 1 October 2008 (i.e., from when the CIN census began). Where a UPN is returned, the aPMR is available, enabling linkage to the NPD education records. For data before April 2005, it is not possible to link education and the CLA datasets.

It is not possible to link social care and education records for children who were only in contact with children's social care services before their UPN was assigned (i.e., for most children who were a child in need or looked after before school age). Previous research has shown that 20% of children who are ever looked after during childhood are only looked after before age 5 (Mc Grath-Lone, Etoori, *et al.*, 2022).

###Children Looked After Return (CLA)

Years available: 2006 to 2022

Information on children looked after by a local authority in England for a period of at least 24 hours (Mc Grath-Lone, Harron, *et al.*, 2016; Department for Education, 2024f, 2024a). The data include information on date and type of placement, use of respite care, and exiting from care, including through adoption. Does not include information on informal fostering arrangements.

6.4.1 Children in Need Census (CIN)

Years available: 2008 to 2022

Has information covering all children who are referred to children's social care services, including information on whether they were assessed and found to be in need (Emmott, Jay and Woodman, 2019; Department for Education, 2023a, 2024e). The CIN census also contains information on children who are subject to a Child Protection Plan (CPP).

6.5 Individualised Learner Record data collection

Years available: 2001 to 2022

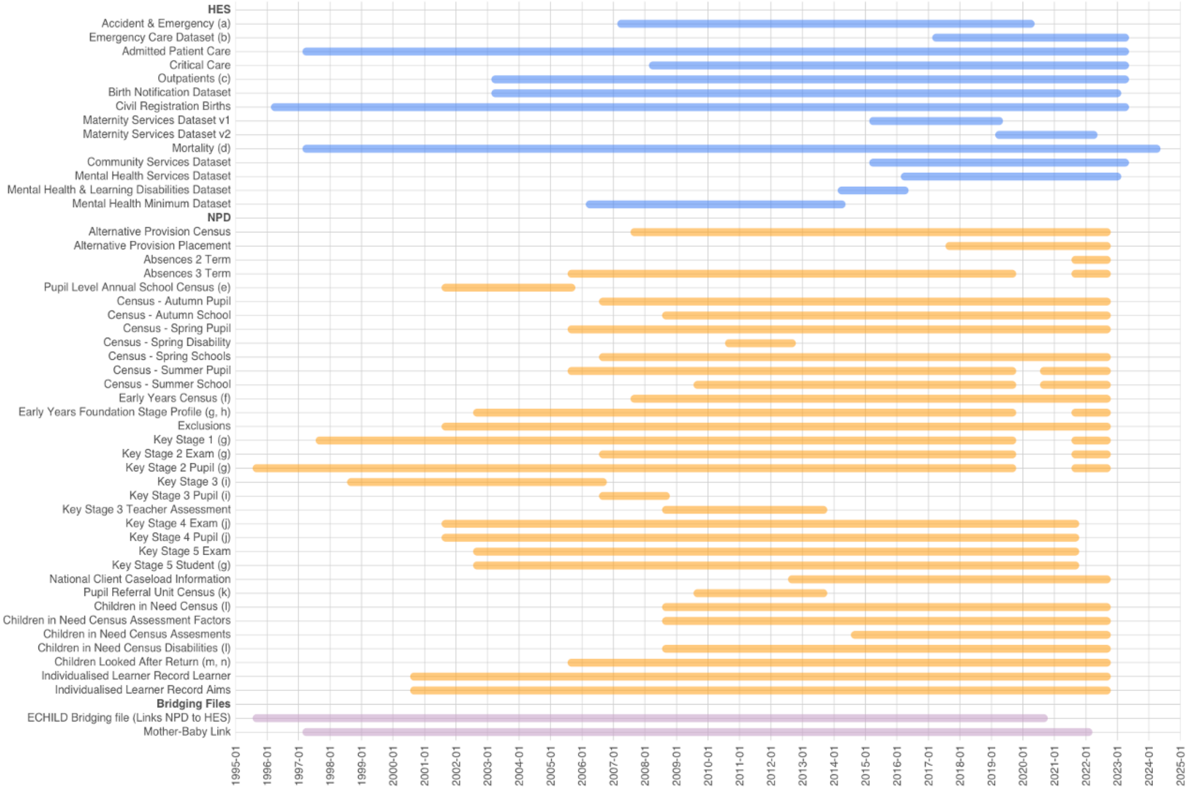
Training providers within the Further Education (FE) and skills sector in England use the 'Individualised Learner Record' (ILR) to collect information about each of the learners in their sector, the learning undertaken, and the learning outcome, e.g., sectors include Adult skills, Community Learning, Skills Bootcamps, 16-19 (excluding Apprenticeships) (Department for Education, 2023c).

ILR data are used to ensure public money distributed through the Education & Skills Funding Agency is being spent in line with government targets, for quality, value for money, planning and supporting future initiatives.

An ILR 'Year' of data, typically runs from 1st August to 31st July and in May 2023.

7 Data coverage (years)

Coverage of the Attribute data currently available within ECHILD (years of follow-up).



Key	Detail
a	Data from 2007/08 to 2011/12 were classified as "experimental" by NHS England.
b	ECDS data from October 2017 to 2019/20 are considered pilot data, with ECDS formally replacing HES A&E in 2020/21.
c	Data from 2003/04 to 2007/08 were classified as "experimental" by NHS England.
d	ONS Mortality data were first linked to HES in January 1998.

Key	Detail
e	From 2001/02 to 2004/05 annual census exists, from 2005/06 onwards census data are termly.
f	The Early Years Census included 3- to 4-year-olds between 2007/08 and 2012/13. From 2013/14 it also includes 2- to 4-year-olds.
g	Not collected in 2019/20 and 2020/21 academic years due to COVID.
h	Partial coverage as between the 2002/03 and 2005/06 academic years, data only on a 10% sample of children.
i	Key Stage 3 assessments ceased after 2012/13.
j	Institutional identifiers not included in 2019/20 and 2020/21 academic years due to restrictions on comparing institutional performance during the pandemic years.
k	PRU Census data were subsumed into the School Census Pupil Level table from 2013/14.
l	CIN: Linkage to NPD is incomplete prior to September 2012.
m	Partial coverage of population as between 01/04/1992 & 31/03/2003, CLA data were only collected for a one-third sample (i.e., children with a day of birth divisible by 3).
n	CLA: Linkage to NPD from 1 April 2005 onwards.

Part II

Linkage

8 Linkage

8.1 Linkage between health and education data in ECHILD

The ECHILD linkage spine is generated by NHS England. NHS England receives identifiers from the DfE (name, date of birth, sex, postcode) and the aPMR. The DfE identifiers are linked with NHS identifiers, and are held separately from any health or education information (NHS England, 2023h). For each matched pair of identifiers from education and health, NHS England attaches a pseudonymised ID called a 'Token Person ID' (TPI) (NHS England, 2024d). The TPI is created specifically for ECHILD and cannot be used to identify individuals. Further detail on the linkage processes can be found in Appendix A Appendix A.

A bridging file, containing the linked aPMR and TPI, can be used to link data for individuals across health and education datasets within ECHILD.

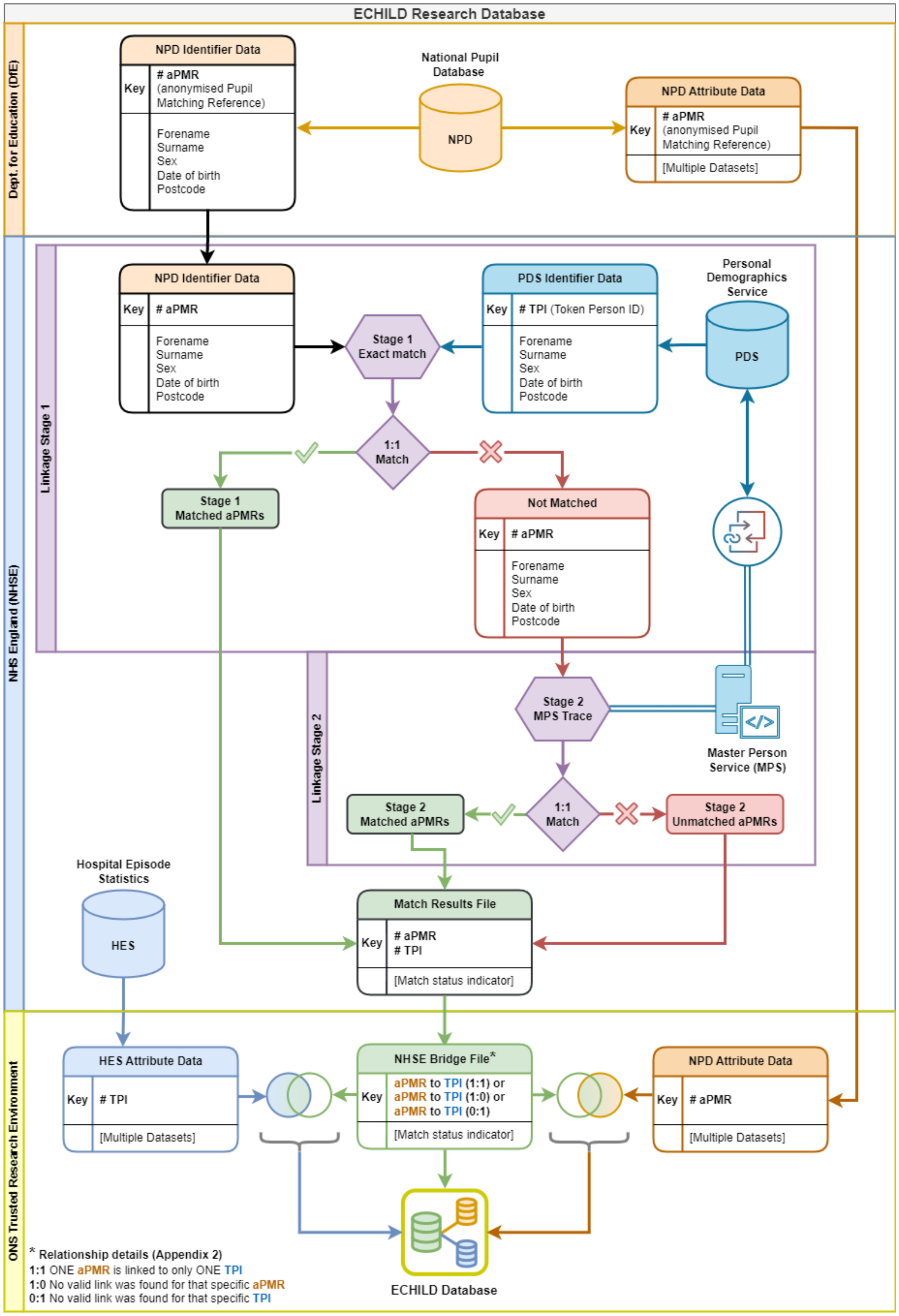

ECHILD Research Database
Dept. for Education (DfE)
NPD Identifier Data
Key
aPMR
(anonymised Pupil Matching Reference)
Forename
Surname
Sex
Date of birth
Postcode
National Pupil Database
NPD
NPD Attribute Data
Key
aPMR
(anonymised Pupil Matching Reference)
[Multiple Datasets]
NHS England (NHSE)
Linkage Stage 1
NPD Identifier Data
Key
aPMR
Forename
Surname
Sex
Date of birth
Postcode
Stage 1
Exact match
PDS Identifier Data
Key
TPI (Token Person ID)
Forename
Surname
Sex
Date of birth
Postcode
Personal Demographics Service
PDS
1:1
Match
Stage 1
Matched aPMRs
Not Matched
Key
aPMR
Forename
Surname
Sex
Date of birth
Postcode
Linkage Stage 2
Stage 2
MPS Trace
Master Person Service (MPS)
Stage 2
Matched aPMRs
1:1
Match
Stage 2
Unmatched aPMRs
Hospital Episode Statistics
HES
Match Results File
Key
aPMR
TPI
[Match status indicator]
ONS Trusted Research Environment
HES Attribute Data
Key
TPI
[Multiple Datasets]
NHSE Bridge File*
Key
aPMR to TPI (1:1) or
aPMR to TPI (1:0) or
aPMR to TPI (0:1)
[Match status indicator]
NPD Attribute Data
Key
aPMR
[Multiple Datasets]
ECHILD Database
* Relationship details (Appendix 2)
1:1 ONE aPMR is linked to only ONE TPI
1:0 No valid link was found for that specific aPMR
0:1 No valid link was found for that specific TPI

Figure 5:

Overview of processes involved in the creation of the ECHILD database.

8.2 Mother-baby linkage in ECHILD

The ECHILD team have created a ‘mother-baby’ link (Feng *et al.*, 2024a), whereby HES delivery and birth records for both mother and baby are linked together using a probabilistic linkage algorithm (Harron *et al.*, 2016). In summary, the probabilistic linkage approach uses ‘indirect identifiers’ (such as gestational age and birth weight) to link together mother-baby dyads. Previous work has demonstrated high linkage rates using this approach (Feng *et al.*, 2024b). The mother-baby linkage means that it is possible within ECHILD to look at how maternal characteristics (including those captured in health, education and social care datasets) are related to child outcomes. It is also possible to identify sibling clusters.

Part III

Strengths & Limitations

9 Strengths

9.1 Includes information related to health, education & social care

HES and NPD are well-established administrative datasets for health, education and social care in England. These datasets act as an evidence base to inform policy: they are used to produce national statistics by government departments and for wider research purposes by the academic community. However, the lack of a common identifier in these administrative datasets has limited the potential for wide-scale analysis across domains. By linking health, education and social care datasets, ECHILD presents a unique and valuable opportunity to explore how children's health affects their education, and how their education affects their health.

9.2 Longitudinal data resource for a whole population-based cohort of children & young people

All children and young people in England who were born between 01/09/1984 to date (updated annually) are eligible for inclusion in ECHILD. Overall, the dataset contains linked health and education records for approximately 20 million individuals. The large sample size and long follow-up period will enable research into long-term outcomes and less common exposures. The ambition is for ECHILD to be updated in the future to include more recent years of data as they become available. This would extend the length of follow-up for cohort members who were born more recently.

9.3 Comprises well-documented administrative datasets

The constituent datasets within ECHILD are well-documented by data owners and the research community. For example, details about how information in the datasets is collected, what variables they contain and how coding has changed over time are readily available to researchers.

9.4 Provides timely access to administrative data for research purposes

Negotiating access with multiple administrative data providers is time-consuming and resource intensive, particularly when it involves the transfer of identifiable information for linkage purposes (Morris, Lanati and Gilbert, 2018). Governance arrangements for the re-use of the de-identified ECHILD data for research purposes via the ONS SRS are now established. This will avoid the need for repeated transfer and use of identifiable information to link HES and NPD data for individual research projects.

10 Limitations

10.1 Administrative data are not collected for research purposes

Administrative datasets are not specifically collected for research purposes, which has implications for the type of research that can be carried out and how research findings are interpreted (Playford *et al.*, 2016). For example, HES data are primarily used for reimbursement of costs and so there may be differences in the frequency and quality of the information that is recorded based on the impact it has on payment. Researchers who intend to carry out secondary analysis of ECHILD must familiarise themselves with the constituent datasets to understand the potential limitations and caveats of their proposed analyses.

10.2 Potential for linkage error

Firstly, there may be errors in the linkage of records within HES (by TPI) or NPD (by aPMR). As previously outlined, TPI and aPMR are derived using linkage algorithms that use various combinations of identifiable information, including name, date of birth, postcode and NHS number or UPN. Secondly, there may be errors in the linkage between HES and NPD that was carried out to create the ECHILD database. Initial evaluation of linkage quality found that approximately 97% of children recorded in NPD matched to a HES record, but that minority ethnic groups and pupils from more disadvantaged neighbourhoods were less likely to be linked (Libuy *et al.*, 2021).

10.3 Constituent datasets in ECHILD have different structures

Both HES and NPD contain individual-level data; however, the structure of the dataset modules varies between (and within) HES and NPD. For example, HES is an episode-level dataset where each row represents a period of continuous care from a consultant, outpatient appointment or A&E attendance, depending on the data module. NPD, CIN and CLA are also episode-level data modules where each row represents a referral to Children's Social Care services (within which there is a significant degree of duplication) or a period of time a child was looked after under a specific legal status and in a specific placement setting, respectively. NPD census modules contain enrolment-level information which means that children who are simultaneously enrolled in more than one educational setting will have multiple rows of information recorded. These differences in data structure mean that researchers will need to carry out substantial dataset manipulation prior to their analyses.

Data from other health providers such as General Practitioners (GPs) or pharmacies are not included.

Part IV

Data Access

11 Five Safes

ECHILD is available to researchers through the Office for National Statistics Secure Research Service (ONS SRS), which follows the Five Safes Framework.

1. Safe People	2. Safe Projects	3. Safe Settings	4. Safe Outputs	5. Safe Data
Accredited Researchers	Ethical & benefits public	Secure technology	Non-identifiable outputs	Use de-identified data

12 Ethics self-assessment

It is important to consider the ethical aspects of any study involving the secondary analysis of de-identified administrative data; although the data are de-identified, no study is entirely free of risk. While completing a Self-Assessment form is not required to apply for access to ECHILD data, it is still recommended. This form helps researchers assess their project's compliance with the ethical principles developed by the National Statistician's Data Ethics Advisory Committee.

13 Access conditions

Researchers will have to demonstrate how their research will benefit the health and well-being of children and young people accessing health, education and social care services. More specifically, projects will have to fall under at least one of the agreed five research purposes described in Chapter 3. We do not have approval for non-UCL PhD students without a substantive contract at an applying institution to access ECHILD.

Researchers will be expected to:

1. State their use of ECHILD in any publication/presentation and acknowledge the ECHILD team in publications and reports, including the following acknowledgements/notes where possible: *We would like to acknowledge the contribution of the wider ECHILD Database support and programme management.*
2. Undertake appropriate Public and Patient Involvement and Engagement (PPIE) activities.
3. Refer to ONS guidance regarding pre-publication, publication and code file clearance.
4. Notify and provide details in writing of all publications to UCL 14 days in advance of publication.
5. Notify and provide draft publications to the Department for Education 14 days in advance of any publication.
6. Include the following data sharing and funding statements in publications as defined by NHS England, the Department for Education, the ONS and ECHILD Database funders:
    a) *The ECHILD Database uses data from the Department for Education (DfE). The DfE does not accept responsibility for any inferences or conclusions derived by the authors.*
    b) *This work uses data provided by patients and collected by the National Health Service as part of their care and support. Source data can also be accessed by researchers by applying to NHS England.*
    c) *We are grateful to the Office for National Statistics (ONS) for providing the Trusted Research Environment for the ECHILD Database. ONS agrees that the figures and descriptions of results in the attached document may be published. This does not imply ONS' acceptance of the validity of the methods used to obtain these figures, or of any analysis of the results.*
    d) *The views in this publication do not necessarily reflect the views of UCL.*
7. Report how data quality issues were addressed.
8. Share their code/script for data processing and analysis.

9. Prepare and send an annual report to the ECHILD team on the benefits of the research undertaken.

14 Process to access ECHILD

Substantive employees, or PhD students with a substantive contract, within UK universities or other research organisation can apply to access ECHILD using the following process:

1. Obtain Full Accredited Researcher status.
2. Familiarise yourself with the ECHILD User Guide and Data Catalogue.
3. Ensure that your institution has Assured Organisational Connectivity certified by the ONS Secure Research Service, or alternatively, check you can access the SafePod Network.
4. Contact the ECHILD team at UCL to discuss feasibility: ich.echild@ucl.ac.uk
5. We recommend completing the UKSA Ethics self-assessment, but it is not compulsory.
6. Complete a project Application Form and a Data Request Form, and submit to the ECHILD team.
   Note: We aim to provide feedback within 2 weeks of receiving a completed application form.
7. Once your application has been approved by UCL, you will submit your application to the ONS PASS system for project accreditation.
   Note: applications take an average of 10 weeks to be processed by ONS.
8. Finally, your organisation will need to sign a **non-negotiable** Data Access Agreement (sublicense) with UCL.

15 Accredited researcher status

To access ECHILD researchers must become fully Accredited Researchers under the Digital Economy Act 2017 (DEA) as this enables researchers to carry out analysis and produce outputs on projects within a Trusted Research Environment such as the ONS Secure Research Service (SRS). To be a fully accredited researcher, individuals must have an undergraduate degree (or higher), including a significant proportion of maths or statistics. Otherwise, they must be able to demonstrate at least three years of quantitative research experience (Office for National Statistics, 2023a).

16 Assured Organisational Connectivity (AOC)

An AOC agreement is an agreement between your organisation and the ONS concerning how your organisation meets the required physical and system security standard to directly allow access to the SRS from your organisation or your home office space. All AOC agreements must be approved by the ONS and it is the responsibility of the researcher to check if an institutional AOC is in place before applying for access to ECHILD.

17 ECHILD cannot be linked to other data sources without further approval

Researchers wishing to link additional datasets to ECHILD will need permission from the data controllers i.e., NHSE and DfE. In addition, the ECHILD team will consider the feasibility of such requests on a case-by-case basis and encourage researchers to contact the team for further discussion.

Part V

References

Reference List

Administrative Data Research UK (2023) ‘ECHILD: Linking children’s health and education data for england’. Available at: https://www.adruk.org/our-work/browse-all-projects/echild-linking-childrens-health-and-education-data-for-england-142/.

Benchimol, E.I. *et al.* (2015) ‘The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement’, *PLoS Med*, 12(10), p. e1001885.

Blackburn, R. *et al.* (2021) ‘Hospital admissions for stress-related presentations among school-aged adolescents during term time versus holidays in england: Weekly time series and retrospective cross-sectional analysis’, *BJPsych Open*, 7(6), p. e215.

Blackburn, R. *et al.* (2022) ‘COVID-19-related school closures and patterns of hospital admissions with stress-related presentations in secondary school-aged adolescents: Weekly time series’, *Br J Psychiatr*, 221(5), pp. 655–657.

Boyd, A. *et al.* (2018) ‘Understanding hospital episode statistics (HES)’. Available at: https://www.closer.ac.uk/wp-content/uploads/CLOSER-resource-understanding-hospital-episode-statistics-2018.pdf.

Clery, A. *et al.* (2024) ‘Can administrative data be used to research health visiting in england? A completeness assessment of the community services dataset’, *International Journal of Population Data Science*, 9(1). doi:10.23889/ijpds.v9i1.2385.

Clinical Practice Research Datalink (2021) ‘Clinical practice research datalink (CPRD): Hospital episode statistics (HES) admitted patient care (APC) data dictionary’. Available at: https://cprd.com/sites/default/files/2022-02/Data_Dictionary_HES_APC.pdf.

CSCDUG (2023) ‘The children’s social care data user group (CSCDUG)’. Available at: https://cscdug.co.uk/.

Department for Education (2015) ‘The national pupil database user guide’. Available at: http://doc.ukdataservice.ac.uk/doc/7627/mrdoc/pdf/7627userguide_2015.pdf.

Department for Education (2017) ‘Exclusions statistics guide’. Available at: https://www.gov.uk/government/publications/exclusions-statistics-guide.

Department for Education (2019) ‘Unique pupil numbers (UPNs)’. Available at: https://www.gov.uk/government/publications/unique-pupil-numbers.

Department for Education (2022) ‘The school census – what you need to know’. Available at: https://educationhub.blog.gov.uk/2022/10/07/the-school-census-what-you-need-to-know/.

Department for Education (2023a) ‘Children in need census: Guide to submitting data’. Available at: https://www.gov.uk/guidance/children-in-need-census.

Department for Education (2023b) ‘Find and explore data in the national pupil database’. Available at: https://find-npd-data.education.gov.uk/.

Department for Education (2023c) ‘Individualised learner record (ILR) technical documents, guidance and requirements’. Available at: https://guidance.submit-learner-data.service.gov.uk/.

Department for Education (2023d) ‘NCCIS management information requirement’. Available at: https://www.gov.uk/government/publications/nccis-management-information-requirement.

Department for Education (2023e) ‘Pupil absence statistics: guide’. Available at: https://www.gov.uk/government/publications/absence-statistics-guide.

Department for Education (2024a) ‘Children looked after return: Guide to submitting data’. Available at: https://www.gov.uk/guidance/children-looked-after-return-guide-to-submitting-data.

Department for Education (2024b) ‘Complete the school census’. Available at: https://www.gov.uk/guidance/complete-the-school-census.

Department for Education (2024c) ‘Early years census’. Available at: https://www.gov.uk/guidance/complete-the-early-years-census.

Department for Education (2024d) ‘Get information about schools’. Available at: https://get-information-schools.service.gov.uk/.

Department for Education (2024e) ‘Statistics: Children in need and child protection’. Available at: https://www.gov.uk/government/collections/statistics-children-in-need.

Department for Education (2024f) ‘Statistics: Looked-after children’. Available at: https://www.gov.uk/government/collections/statistics-looked-after-children.

ECHILD (2023) ‘The ECHILD project’. Available at: https://www.echild.ac.uk.

Emmott, E.H., Jay, M.A. and Woodman, J. (2019) ‘Cohort profile: Children in need census

(CIN) records of children referred for social care support in england', *BMJ Open*, 9(2), p. e023771.

Etoori, D. *et al.* (2022) 'Reductions in hospital care among clinically vulnerable children aged 0–4 years during the COVID-19 pandemic', *Arch Dis Child*, 107, p. e29.

Feng, Q. *et al.* (2024a) 'Data resource profile: A national linked mother-baby cohort of health, education and social care data in england (ECHILD-MB)', *Int J Epidemiol*, 53(3). doi:/10.1093/ije/dyae065.

Feng, Q. *et al.* (2024b) 'Data resource profile: ECHILD only-children and siblings (ECHILD-oCSib): A national cohort of linked health, education and social care data on mothers and children in england', *IJPDS*, 8(6). doi:/10.23889/ijpds.v8i6.2392.

Gilbert, R. *et al.* (2017) 'GUILD: GUidance for Information about Linking Data sets', *J Pub Health*, 40(1), pp. 191–198.

Green, F. *et al.* (2017) 'Who chooses private schooling in britain and why?' Available at: https://www.llakes.ac.uk/wp-content/uploads/2021/03/RP-62.-Green-Anders-Henderson-Henseke.pdf.

Harron, K. *et al.* (2016) 'Linking data for mothers and babies in de-identified electronic health data', *PLoS One*, 11(10), p. e0164667.

Harron, K. *et al.* (2017) 'A guide to evaluating linkage quality for the analysis of linked data', *Int J Epidemiol*, 46, pp. 1699–1710.

Health Data Research (2023a) 'Innovation gateway: Hospital episode statistics accident and emergency (HES a&e)'. Available at: https://web.www.healthdatagateway.org/dataset/101a7d68-1675-4b47-bfac-dde7b7b57877.

Health Data Research (2023b) 'Innovation gateway: Hospital episode statistics admitted patient care (HES APC)'. Available at: https://web.www.healthdatagateway.org/dataset/6599230a-df54-4615-937c-d724d239491f.

Health Data Research (2023c) 'Innovation gateway: Hospital episode statistics critical care (HES CC)'. Available at: https://web.www.healthdatagateway.org/dataset/d0eeba22-bc3f-4c40-be69-a99f916897a2.

Health Data Research (2023d) 'Innovation gateway: Hospital episode statistics outpatients'. Available at: https://web.www.healthdatagateway.org/dataset/d4362d41-0838-4f41-bba3-c3268312c444.

Health Data Research (2023e) 'Innovation gateway: Maternity services data set (MSDS)'. Available at: https://web.www.healthdatagateway.org/dataset/0681a1e9-3778-4f43-88b3-6540515404b6.

Herbert, A. *et al.* (2017) 'Data resource profile: Hospital episode statistics admitted patient care', *Int J Epidemiol*, 46(4), pp. 1093–1093i.

Jay, M.A., Mc Grath-Lone, L. and Gilbert, R. (2019) 'Data resource: The national pupil database (NPD)', *Int J Pop Data Sci*, 4(1), p. 08.

Libuy, N. *et al.* (2021) 'Linking education and hospital data in england: Linkage process and quality', *Int J Pop Data Sci*, 6(1), p. 13.

Libuy, N. *et al.* (2022) 'Gestational age at birth, chronic conditions and school outcomes: A population-based data linkage study of children born in england', *Int J Epidemiol*, 52(1), pp. 132–143.

Mc Grath-Lone, L., Dearden, L., *et al.* (2016) 'Changes in first entry to out-of-home care from 1992 to 2012 among children in england', *Child Abuse & Neglect*, 51, pp. 163–171.

Mc Grath-Lone, L., Harron, K., *et al.* (2016) 'Data resource profile: Children looked after return (CLA)', *Int J Epidemiol*, 45(3), pp. 716–717f.

Mc Grath-Lone, L., Libuy, N., Harron, K., *et al.* (2021) 'Data resource profile: The education and child health insights from linked data (ECHILD) database', *Int J Epidemiol*, 51(1), pp. 17–17f.

Mc Grath-Lone, L., Libuy, N., Etoori, D., *et al.* (2021) 'Ethnic bias in data linkage', *Lancet Digital Health*, 3(6), p. e339.

Mc Grath-Lone, L., Etoori, D., *et al.* (2022) 'Changes in adolescents' planned hospital care during the COVID-19 pandemic: Analysis of linked administrative data', *Arch Dis Chil*, 107(10), p. e29.

Mc Grath-Lone, L., Jay, M.A., *et al.* (2022) 'What makes administrative data "research-ready"? A systematic review and thematic analysis of published literature', *Int J Pop Data Sci*, 7(1), p. 1718.

Morris, H., Lanati, S. and Gilbert, R. (2018) 'Challenges of administrative data linkages: Experiences of administrative data research centre for england (ADRC-E) researchers', *Int J Pop Data Sci*, 3(2), p. 097.

Nguyen, V.G. *et al.* (2022) 'SEN support from the start of school and its impact on unplanned hospital utilisation in children with cleft lip and palate: A demonstration target trial emulation protocol using ECHILD'. Available at: https://www.medrxiv.org/content/10.1101/2022.04.01.22273280v1.

NHS England (2023a) 'About the maternity services data set (MSDS)'. Available at: https://digital.nhs.uk/data-and-information/data-collections-and-data-sets/data-sets/maternity-services-data-set/about-the-maternity-services-data-set.

NHS England (2023b) 'Birth notification process'. Available at: https://digital.nhs.uk/services/national-care-records-service/birth-notification-process.

NHS England (2023c) 'ECDS guidance and documents'. Available at: https://digital.nhs.uk/data-and-information/data-collections-and-data-sets/data-sets/emergency-care-data-set-ecds/ecds-guidance.

NHS England (2023d) 'Emergency care data set (ECDS)'. Available at: https://digital.nhs.uk/data-and-information/data-collections-and-data-sets/data-sets/emergency-care-data-set-ecds.

NHS England (2023e) 'Hospital accident & emergency activity'. Available at: https://digital.nhs.uk/data-and-information/publications/statistical/hospital-accident--emergency-activity.

NHS England (2023f) 'Hospital episode statistics data dictionary'. Available at: https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/hospital-episode-statistics/hospital-episode-statistics-data-dictionary.

NHS England (2023g) 'Hospital episode statistics (HES)'. Available at: https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/hospital-episode-statistics.

NHS England (2023h) 'Hospital episode statistics (HES) data changes in 2021'. Available at: https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/hospital-episode-statistics/hospital-episode-statistics-data-changes-in-2021.

NHS England (2023i) 'Hospital outpatient activity'. Available at: https://digital.nhs.uk/data-and-information/publications/statistical/hospital-outpatient-activity.

NHS England (2023j) 'Introduction to community services data set (CSDS)'. Available at: https://digital.nhs.uk/data-and-information/data-collections-and-data-sets/data-sets/community-services-data-set/guidance/introduction.

NHS England (2023k) 'Master person service (MPS)'. Available at: https://digital.nhs.uk/services/demographics/master-person-service.

NHS England (2023l) 'Mental health services data set (MHSDS)'. Available at:

https://digital.nhs.uk/data-and-information/data-collections-and-data-sets/data-sets/mental-health-services-data-set/about.

NHS England (2024a) ‘Hospital admitted patient care activity’. Available at: https://digital.nhs.uk/data-and-information/publications/statistical/hospital-admitted-patient-care-activity.

NHS England (2024b) ‘Hospital adult critical care activity’. Available at: https://digital.nhs.uk/data-and-information/publications/statistical/hospital-adult-critical-care-activity.

NHS England (2024c) ‘Mental health data (MHMDS, MHLDDS, MHSDS)’. Available at: https://digital.nhs.uk/services/data-access-request-service-dars/dars-products-and-services/data-set-catalogue/mental-health-data.

NHS England (2024d) ‘Methodological changes’. Available at: https://digital.nhs.uk/data-and-information/find-data-and-publications/statement-of-administrative-sources/methodological-changes.

NHS Primary Care Support England (2023) ‘Adoptions and gender reassignment’. Available at: https://pcse.england.nhs.uk/help/patient-registrations/adoption-and-gender-re-assignment-processes/.

Office for National Statistics (2023a) ‘Access the data securely’. Available at: https://www.ons.gov.uk/aboutus/whatwedo/statistics/requestingstatistics/secureresearchservice/accessthedatasecurely#assured-organisational-connectivity-aoc.

Office for National Statistics (2023b) ‘Secure research service’. Available at: https://www.ons.gov.uk/aboutus/whatwedo/statistics/requestingstatistics/secureresearchservice.

Office for National Statistics (2024) ‘User guide to birth statistics’. Available at: https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/livebirths/methodologies/userguidetobirthstatistics.

Playford, C.J. *et al.* (2016) ‘Administrative social science data: The challenge of reproducible research’, *Big Data & Society*, 3, p. doi:10.1177/2053951716684143.

Scourfield, J. *et al.* (2019) ‘Overview of administrative data on children’s social care in england’. Available at: https://whatworks-csc.org.uk/wp-content/uploads/WWCSC_Overview_of_administrative_data_Oct_2019.pdf.

UK Statistics Authority (2023) ‘Ethics self-assessment tool’. Available at: https://uksa.statisticsauthority.gov.uk/the-authority-board/committees/national-statisticians-advisory-committees-and-panels/national-statisticians-data-ethics-advisory-committee/ethics-self-assessment-tool/.

University of Bristol (2023) ‘National pupil database user group’. Available at: http://www.bristol.ac.uk/cmpo/npd-user-group/.

Part VI

About

18 Acronyms

Acronym	Definition
A&E	Accident & Emergency
ADRUK	Administrative Data Research UK
AOC	Assured Organisational Connectivity
AP	Alternative Provision
APC	Admitted Patient Care
aPMR	anonymised Pupil Matching Reference
CC	Critical Care
CIN	Children in Need
CLA	Children Looked After
CPP	Child Protection Plans
CSDS	Community Services Data Set
DEA	Digital Economy Act
DfE	Department for Education
ECHILD	Education & Child Health Insights from Linked Data
EHC	Education, Health & Care
EYFSP	Early Years Foundation Stage Profile
FAE	Finished Admission Episode
FCE	Finished Consultant Episode
FE	Finished Episode
FSM	Free School Meals
GIAS	Get Information About Schools
GP	General Practitioner
HES	Hospital Episode Statistics
ICD	International Classification of Diseases
ILR	Individualised Learner Record
KS	Key Stage
MHLDDS	Mental Health and Learning Disabilities Data set
MHMDS	Mental Health Minimum Data Set
MHSDS	Mental Health Services Data Set
MPS	Master Person Service
MSDS	Maternity Services Data Set
NCCIS	National Client Caseload Information System
NEET	Not in Education, Employment or Training

Acronym	Definition
NHSE	National Health Service England
NIHR	National Institute for Health and Care Research
NPD	National Pupil Database
ONS	Office for National Statistics
OP	Outpatient
OPCS	Office of Population Censuses and Surveys
PPIE	Public and Patient Involvement and Engagement
PRU	Pupil Referral Unit
PVI	Private, Voluntary and Independent
RAP	Research Accreditation Panel
SEN	Special Educational Needs
SRS	Secure Research Service
TPI	Token Person ID
UPN	Unique Pupil Number

19 Versions

The website version of the ECHILD User Guide is the canonical source of the ECHILD User Guide. A pdf version of the contents of this website is automatically generated and made available without warranty.

Version	Date	Details
2.1.2 onwards	See release history	See release notes.
2.1.1	27 July 2024	First web version based on version 2.0, ported by Tony Stone, Matthew Jay and Farzan Ramzan.
2.0	22 June 2023	Education and Child Health Insights from Linked Data (ECHILD): An Introductory Guide for Researchers
1.1.2	5 March 2021	The Education and Child Health Insights from Linked Data (ECHILD) Database: An Introductory Guide for Researchers

20 Disclaimer

Although every effort has been made to provide complete and accurate information at the time of publication (June 2023), the authors make no warranties, express or implied, or representations as to the accuracy of content in this guide. The authors assume no liability or responsibility for any error or omissions in the information contained in the guide.

21 Acknowledgements

ECHILD is led by University College London Great Ormond Street Institute of Child Health in collaboration with the London School of Hygiene & Tropical Medicine and the Institute for Fiscal Studies, in partnership with the Department of Health and Social Care and the Department for Education, working with NHS England and the Office for National Statistics.

We thank the following individuals for their valuable contributions to the project. From NHS England, we are grateful to Jodie Taylor-Brown, Ian Goodwin, Garry Coleman, Richard Caulton, Catherine Day, Terry Service, and Narissa Leyland. From the Department for Education (DfE), we thank Gary Connell, Harriet Fearn, Kirsty Knox, and Neide Phillips. We are also grateful to Bill South and Alan Cotterill from the Office for National Statistics (ONS) for providing the Trustworthy Research Environment for ECHILD. In addition, we thank Yoli Bartlett, Zuzanna Domaradzka, Cameron Kelly, Tudor Vilcan, Rachel Orchard, and all those in the ONS Secure Research Service Support, Applications, and Operations teams for their vital support.

We thank all the children, young people, parents and carers who contributed to the ECHILD project. We also gratefully acknowledge all children and families whose de-identified data form this research database.

This user guide describes the ECHILD database which uses data from the DfE, NHS England and ONS. The DfE, NHS England and ONS do not accept responsibility for any inferences or conclusions derived by the authors.

This research benefits from and contributes to research conducted by the NIHR Children and Families Policy Research Unit but was not commissioned by the National Institute for Health Research (NIHR) Policy Research Programme. The views expressed herein are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care.

22 Authors

22.1 UCL Great Ormond Street Institute of Child Health

- Dr Farzan Ramzan
- Dr Louise Mc Grath-Lone (now at UCL Institute of Education)
- Dr Ruth Blackburn
- Prof Ruth Gilbert
- Dr Matthew Jay
- Dr Kate Lewis
- Mr Matthew Lilliman
- Dr Vincent Nguyen
- Dr Milagros Ruiz Nishiki
- Mr Tony Stone
- Prof Katie Harron

23 Contributors

23.1 UCL Great Ormond Street Institute of Child Health

- Prof Pia Hardelid
- Dr Nicolás Libuy
- Dr Ania Zylbersztejn

23.2 Institute for Fiscal Studies

- Ms Christine Farquharson
- Mr Imran Tahir

24 Funding

This work is funded via Administrative Data Research UK (ADR UK), an investment by the Economic and Social Research Council (part of UK Research and Innovation), through the following grants:

- ES/V000977/1
- ES/X000427/1
- ES/X003663/1

25 Contact details

For further information, comments or queries about this publication:

Website	Email	X (formerly Twitter)
www.echild.ac.uk	ich.echild@ucl.ac.uk	@ucl_echild

A Linkage process

This section provides an overview of the linkage process (see figure under ECHILD database structure) used to create the ECHILD database.

A.1 Data sources

A.1.1 The Department for Education's National Pupil Database

The Department for Education (DfE) collates and manages data on school children in England in a resource known as the National Pupil Database (NPD) (Jay, Mc Grath-Lone and Gilbert, 2019). For the purposes of the ECHILD linkage, salient points are described below.

Within NPD, each record is associated with natural identifiers (forename, surname, gender, date of birth, postcode) relating to the pupil's details as known by the submitting organisation at the time the data were submitted. DfE additionally assigns each record in NPD an anonymised Pupil Matching Reference (aPMR). An aPMR is an identifier which is not in itself meaningful: it does not reveal the identity of the pupil. aPMRs are assigned to all records in NPD such that one aPMR should represent one pupil and each pupil should have only one aPMR. This allows records for each pupil to be identified across NPD, both between different datasets and over time, without revealing their identity. As a result, NPD contains a longitudinal record of pupils' names, addresses, and (potentially) genders over time.

A.1.2 NHS England's Personal Demographics Service and Master Person Service

NHS England operates the Personal Demographics Service (PDS), a national electronic database of demographic data for patients accessing care in England or services funded by the NHS in England. PDS contains natural identifiers (forename, surname, gender, date of birth, postcode) and NHS number. Each person is assigned a distinct NHS number at birth if born in England, or at the first time of accessing NHS services in England if not otherwise registered. PDS holds a longitudinal record of name and address changes made to NHS services in England over time. This means there may be many records for each person in PDS but all records for a person should be assigned the same NHS number (with some exceptions, see (NHS Primary Care Support England, 2023)).

The Master Person Service (MPS), managed by NHS England (NHS England, 2023k), takes a record of natural identifiers and attempts to link this to a PDS record allowing for some errors and missingness in the recording of the natural identifiers. If this fails and the natural identifiers are sufficiently complete, MPS attempts to link records against a secondary store (MPS bucket) of natural identifiers of persons who previously had contact with the NHS in England and do not have an NHS number, these persons are assigned an “MPS ID”.

The MPS returns a “Person ID” using either:

1. NHS number, if a valid link is found in PDS; or,
2. MPS ID if no link is found in PDS but a valid link is found in the MPS bucket; or,
3. No value if no link is found in either PDS or the MPS bucket.

The Person ID is then encrypted to generate a Token Person ID (TPI), which is not meaningful and does not reveal the person’s identity.

Person IDs, enabling the assignment of Token Person IDs, are also recorded throughout NHS England’s standard data collections, including Hospital Episode Statistics datasets, Emergency Care Datasets, Mental Health Services Datasets, and Community Services Datasets. The “Person ID” is the only routine means of identifying records belonging to the same patient amongst data held by NHS England.

A.2 Linking DfE NPD aPMRs to NHS England TPIs

For the purposes of the ECHILD linkage, the following simplifying assumptions were made:

1. Each aPMR represents at most one *real* person within NPD;
2. Each TPI represents precisely one *real* person within NHS England data collections;
3. Each *real* person represented within NHS England data collections has precisely one TPI.

Essentially, whilst we assume TPIs are perfectly allocated, we only require that aPMRs are not shared. That is, the same *real* person **is** permitted to have more than one aPMR.

This linkage task resulted in “N-to-one” links between aPMRs and TPIs: each aPMR linked to at most one TPI but a TPI may have linked to more than one aPMR. However, the vast majority of links made were “1-to-1” (99.4% of linked TPIs linked to only one aPMR).

DfE supplied a “linkage dataset” to NHS England, comprising the natural identifiers (forename, surname, gender, date of birth, postcode) and an aPMR for each record in its NPD.

A.2.1 Linkage Stage 1: Exact link

An initial, simple, linkage stage was used to avoid over-burdening the more resource-intensive MPS trace. Each valid record (e.g., no blank entries) in the linkage dataset was compared to all records in a prepared extract from PDS. A sequential, two part, approach was used. Only records belonging to aPMRs not linked after Part A progressed to Part B. In both parts, an aPMR was considered *linked* only if all of its associated linkage records were linked to at most one TPI and at least one record was *linked* to a TPI.

A.2.1.1 Part A: Full forename

A record was deemed *linked* if compared records had equal (non-empty) values across all of the following fields:

1. Full forename;
2. Full surname;
3. Full date of birth;
4. Full postcode; and,
5. Gender.

A.2.1.2 Part B: Partial forename

A record was deemed *linked* if compared records had equal (non-empty) values across all of the following fields:

1. First four characters of forename;
2. Full surname;
3. Full date of birth;
4. Full postcode; and,
5. Gender.

A.2.2 Linkage Stage 2: MPS Trace

Only records belonging to aPMRs not linked after Linkage Stage 1 were submitted to MPS. Again, an aPMR was considered *linked* only if all of its associated linkage records were linked to at most one TPI and at least one record was *linked* to a TPI.

A.2.3 Linkage results

DfE provided 430M records, comprising sets of identifiers, covering 22.8M aPMRs. NHS England managed to link 22.6M (99.1%) aPMRs to a TPI. 20.1M (88.2%) aPMRs were linked at Stage 1, Part A. A further 1.3M (5.7%) aPMRs were linked at Stage 1, Part B. Stage 2 resulted in the linkage of a further 1.2M (5.3%) aPMRs. Less than 200,000 aPMRs remained unlinked at the end of Stage 2: ~160,000 were not resolvable to any TPI; ~30,000 resolved to more than one TPI and (due to the simplifying assumptions) these were considered invalid and so no link was recorded for any of these aPMRs.

A.2.4 Application of NHS National Data Opt-Outs

The ECHILD Research Database team wished to enable potential participants to opt-out from their data being held within the ECHILD Research Database. Our data suppliers indicated that the only means to (partially) operationalise this was through the non-provision of data held by NHS England relating to participants with a current (at the date of data preparation) NHS National Data Opt-Out (NDOO). This included removing any indication of an identified *link* between the aPMRs and TPIs for participants with a NDOO. It was, however, not possible to exclude the DfE supplied data relating to these participants. 1.3M (5.8%) of the 22.6M aPMRs for which NHS England had identified a link to a single TPI related to individuals with a registered NDOO as at the date of generation of the data extract.

A.2.5 Linkage Outputs: Pseudonymised bridging file

NHS England produced a pseudonymised bridging file consisting of all aPMRs in the DfE-supplied linkage dataset and their linked TPI (excluding those removed due to the presence of a NHS National Data Opt-Out). All aPMRs that were not *linked* after Linkage Stage 2, or which related to a participant with a NHS National Data Opt-Out, were omitted from this file.