Linking education and hospital data in England: linkage process and quality

Nicolás Libuy¹,*, Katie Harron¹,², Ruth Gilbert¹,², Richard Caulton³, Ellen Cameron³, and Ruth Blackburn¹

Abstract

Introduction
Linkage of administrative data for universal state education and National Health Service (NHS) hospital care would enable research into the inter-relationships between education and health for all children in England.

Objectives
We aim to describe the linkage process and evaluate the quality of linkage of four one-year birth cohorts within the National Pupil Database (NPD) and Hospital Episode Statistics (HES).

Methods
We used multi-step deterministic linkage algorithms to link longitudinal records from state schools to the chronology of records in the NHS Personal Demographics Service (PDS; linkage stage 1), and HES (linkage stage 2). We calculated linkage rates and compared pupil characteristics in linked and unlinked samples for each stage of linkage and each cohort (1990/91, 1996/97, 1999/00, and 2004/05).

Results
Of the 2,287,671 pupil records, 2,174,601 (95%) linked to HES. Linkage rates improved over time (92% in 1990/91 to 99% in 2004/05). Ethnic minority pupils and those living in more deprived areas were less likely to be matched to hospital records, but differences in pupil characteristics between linked and unlinked samples were moderate to small.

Conclusion
We linked nearly all pupils to at least one hospital record. The high coverage of the linkage represents a unique opportunity for wide-scale analyses across the domains of health and education. However, missed links disproportionately affected ethnic minorities or those living in the poorest neighbourhoods: selection bias could be mitigated by increasing the quality and completeness of identifiers recorded in administrative data or the application of statistical methods that account for missed links.

Keywords
record linkage; linkage error; bias; hospital records; educational records; data linkage; administrative data

Highlights
- Longitudinal administrative records for all children attending state school and acute hospital services in England have been used for research for more than two decades, but lack of a shared unique identifier has limited scope for linkage between these databases.
- We applied multi-step deterministic linkage algorithms to 4 one-year cohorts of children born 1 September-31 August in 1990/91, 1996/97, 1999/00 and 2004/05. In stage 1, full names, date of birth, and postcode histories from education data in the National Pupil Database were linked to the NHS Personal Demographic Service. In stage 2, NHS number, postcode, date of birth and sex were linked to hospital records in Hospital Episode Statistics.
- Between 92% and 99% of school pupils linked to at least one hospital record. Ethnic minority pupils and pupils who were living in the most deprived areas were least likely to link. Ethnic minority pupils were less likely than white children to link at the first step in both algorithms.
- Bias due to linkage errors could lead to an underestimate of the health needs in disadvantaged groups. Improved data quality, more sensitive linkage algorithms, and/or statistical methods that account for missed links in analyses, should be considered to reduce linkage bias.

*Corresponding Author:
Email Address: nicolas.libuy.16@ucl.ac.uk (Nicolás Libuy)
Introduction

Administrative data have been routinely collected for more than two decades in England from schools and hospitals by the Department for Education (DfE) and National Health Service (NHS) Digital respectively [1, 2]. These data collections have been used to monitor service provision and costs, and longitudinal linkage has made them powerful resources for national research [3–7]. Despite evidence from other countries of the value of linking education and health data to inform policy and practice [8–14], these databases have not previously been linked for children in England because they do not share a unique identifier. Linkage between these datasets can only be done using confidential, personal identifiers such as full names, postcodes, date of birth and sex, thereby creating technical and governance challenges.

Linkage error could significantly undermine the real-world benefits for policy if certain groups, such as those with a foreign name structure, are less likely to link than others [15]. For example, missed links could lead to undercounting of adverse health or education outcomes for these groups, and in turn, under-provision of services. Evidence on linkage error can help data providers to improve the quality of identifiers or to develop more effective linkage algorithms. Evidence on differences in the characteristics between groups who link or not can be used by researchers to account for linkage bias in analyses [16].

We describe the methods used to link education data from the National Pupil Database (NPD) to hospital data for children in England (Hospital Episode Statistics; HES) [1, 2]. Our goal was to create de-identified, linked cohorts of pupils’ longitudinal records of education and hospital events over the childhood years. We also evaluated associations between child characteristics and linkage error in order to understand the implications of these errors for analysis. Our evaluation is based on 2.2 million children in England born in four one-year cohorts in 1990/91, 1996/97, 1999/00 and 2004/05. These cohorts reflect age and time periods when identifier quality, and hence linkage quality, is likely to differ due to data collection and system changes. This paper is relevant to users of The Education and Child Health Insights from Linked Data (ECHILD) database, which will be available from Spring 2022 and combines education, social care and hospital data for all children in England born from 1995 [1, 2, 17]. The findings are also relevant more generally to data linkages that lack a unique, high-quality identifier.

Methods

Study design and population

Governance permissions and data flows for the linkage followed the separation principle [16], whereby identifiers such as names and postcodes were kept separate at all times from attribute data (records from school or hospital records). Figure 1 shows the flow of identifiers and a pseudo-identifier (the anonymised Pupil Matching Reference, aPMR) from the Department for Education to NHS Digital. Separately, education attribute data flowed from the Department for Education to the Office of National Statistics Secure Research Service (ONS SRS). A two-stage linkage process was used to link NPD to HES. Stage 1 linked NPD to the Personal Demographic Service (PDS), which contains all individuals with an NHS number, and stage 2 linked NPD-PDS linked data to HES. At the first stage of linkage (step C in Figure 1), NHS Digital linkers had access only to the identifiers (date of birth, sex, and histories of forenames, surnames and postcodes) but no attribute data. At the second stage of linkage (step D), NHS Digital used the NHS number, date of birth, sex and postcode to link to HES data. The linkage step, pseudonymised HESID and anonymised PMR were transferred (step E) and merged with a University College London (UCL) held extract of HES within the UCL Data Safe Haven (DSH) (step F). Linked HES-PMR records were ultimately transferred to the ONS SRS (step G).

The study population consisted of four cohorts of children born between 1 September and 31 August in the academic years of 1990/91, 1996/97, 1999/00, and 2004/05 (Figure 2). These cohorts were defined separately in NPD and HES, so that linkage created three comparison groups for each of the four cohorts: linked NPD-HES, unlinked NPD, and unlinked HES records. We compared pupil characteristics in the linked and unlinked NPD cohorts at each stage of each linkage process. We used NPD as the inception cohort, as state school is a universal service attended at some point in the school years by at least 95% of all children [2, 18]. On the other hand, not all children attend hospital, unless they were young enough for their birth to be recorded in HES (1997 onwards).

Figure 2 shows that whether a pupil is expected to link to a HES record or not is affected by the start date of the PDS, the NPD and the subsets of HES data. Pupils born in 1990/91 were expected to have the lowest proportion of records in NPD that linked to HES (i.e. linkage rate). These children only appeared in NPD at the first school census collection in 2001/02 at age 10. Their names and postcodes captured each year in NPD from 2001/02 until leaving state school in 2009/10 or earlier, would be linked to names and postcode details recorded prospectively from General Practitioner (GP) registrations and hospital contacts on the PDS from 2004 onwards. These children could link to HES admission records from 1997 onwards (age 6 years), outpatients from age 12, or accident and emergency department from age 16.

Whilst it was expected that most children would have contact with hospital at some point during childhood or adolescence, we did not anticipate complete overlap between the two datasets. We expected children born in 2004/05 to have the best linkage rates of the four cohorts (and for linkage quality to remain constant or improve for subsequent cohorts). Firstly, 97% of children born in England would be expected to have their birth recorded in HES and in PDS [19]. Secondly, their linkage to subsequent health records should be more accurate than earlier cohorts due to immediate allocation of adverse health or education outcomes for these groups, and hence linkage quality to remain constant or improve for subsequent cohorts. Firstly, 97% of children born in England would be expected to have their birth recorded in HES and in PDS [19]. Secondly, their linkage to subsequent health records should be more accurate than earlier cohorts due to immediate allocation of adverse health or education outcomes for these groups, and hence linkage quality to remain constant or improve for subsequent cohorts.
Figure 1: Data flow and linkage process for linkage between the national pupil database, the personal demographic service and hospital episode statistics

Figure 2: Lexis diagram to show year of age of each cohort (y axis) and start year of each dataset (x axis)

Notes: NHS Digital sent two Linkage bridging files to UCL DSH. Details are described in Supplementary Appendix 1. Dark shading indicates de-identified and light shading identified data. NPD = National Pupil Dataset; PDS = Personal Demographics Service; HES = Hospital Episode Statistics; NHS = National Health Service; ONS SRS = Office for National Statistics Secure Research Service; UCL = University College London.

Notes: See details in Supplementary Figure 1 and Supplementary Table 1 in the Supplementary Appendices 2 and 3.
National pupil database (NPD)

NPD contains pupil-level information on all children and adolescents attending state-funded schools in England, capturing information on attainment tests, absences, exclusions and alternative provision (details in Supplementary Figure 1 of Supplementary Appendix 2) [2]. The school census collects information each term on pupils enrolled and updates of the pupil’s name, address, and postcode. We used identifiers recorded in the Spring census (submitted in February) for linkage as this is the definitive entry for the year (i.e. for school year 2001/2). Pupil records are linked across years and between NPD modules using a pseudo-identifier called the anonymised Pupil Matching Reference (aPMR).

Hospital episode statistics (HES)

HES is an episode level administrative database that covers all admissions (day case and overnight) to the National Health Service (NHS) hospitals in England [1], as well as all attendances at the accident and emergency departments (from 2007/8) and outpatient appointments (from 2003/4). From January 1998 onwards, HES has been routinely linked to ONS death registration records [21]. Supplementary Table 1 in Supplementary Appendix 3 describes data availability in HES. For researchers using de-identified attribute data from HES, episodes of care relating to a patient can be linked over time or between datasets using a study-specific pseudonymised patient identifier generated by NHD Digital – HESID [22].

Personal demographics service (PDS)

PDS is a national electronic database that contains the chronology of demographic information, including sex, name and address, for all individuals in England with an NHS number. Introduced in June 2004, as part of The National Programme for IT, the PDS was developed to integrate management of patient demographic information across NHS services in England. PDS replaced the NHS Central Register (CHRIS); the demographic functions of the National Health Applications and Infrastructure Services (NHAIS); the NHS Strategic Tracing Service (NSTS); and the NHS Number for Babies (NN4B) [23]. Current identifiers from these databases were transferred into PDS in 2004. The patient demographic details on the PDS data can be updated by NHS care providers when a person uses an NHS service, including GP surgeries, inpatient or outpatient appointments [24, 25]. The accuracy and quality of PDS data is assured by staff at the PDS National Back Office (NBO) in NHS Digital [26].

Linkage

Linkage process

Figure 1 shows two stages of linkage. Stage 1 involved transfer of a linkage file containing full name and postcode histories and other identifiers (Table 1) from the Department for Education to NHS Digital for linkage to the PDS. Extracts from NPD and PDS listed multiple identifiers for each individual together with the date interval when the identifier was recorded (details in Supplementary Appendix 4). To link the NPD linkage file and PDS, we relied on a deterministic linkage algorithm comprising 8 steps, shown in Table 2. These steps were designed to identify records that have high levels of agreement across names, date of birth, sex and postcode, and to resolve inconsistencies between records belonging to the same pupil.

Besides considering the 8 steps in Table 2, a further restriction was that a linked pair of records needed to have identifiers within the same academic year in PDS and in NPD (details in supplementary Appendix 4). All eight steps of the algorithm were run for each school year (September

Table 1: Availability of personal identifiers in the national pupil database, personal demographic service and hospital episode statistics

| Linkage identifiers                              | DfE | NHSD | HES |
|--------------------------------------------------|-----|------|-----|
| First name(s)                                    | ✓   | ✓    |     |
| Surname(s)                                       | ✓   | ✓    |     |
| Date of birth (e.g. 23/02/1988)                   | ✓   | ✓    | ✓   |
| Sex                                              | ✓   | ✓    | ✓   |
| NHS number                                       | ✓   | ✓    | ✓   |
| Residence postcodes*                             | ✓   | ✓    | ✓   |
| Residence postcodes dates**                      | ✓   | ✓    | ✓   |
| Anonymised Pupil Matching Reference (aPMR)       | ✓   | ✓    |     |
| UCL HESID                                        | ✓   | ✓    | ✓   |

Notes: * Full postcodes (e.g. LS0 0AA) were available in NPD and PDS. For records in NPD a list of postcodes was available over the academic years. For a specific patient’s NHS number in PDS, a list of postcodes was available over time.
** Dates referring to changes in patient’s postcodes over time were available in PDS. Similarly, dates referring to postcodes in academic years were available in NPD. UCL HESID: is a unique and pseudonymised patient-level identifier that can be used to link patient-level information over time and across different modules of the UCL HES extracts.

aPMR: anonymised Pupil Matching Reference is a nationally unique and anonymised child-level identifier that can be used to link pupil-level information over time and across different modules of NPD.
Table 2: Linkage stage 1: 8 step deterministic algorithm for linking the national pupil database to the personal demographic service

| Step | First name | Surname | Date of birth | Sex | Postcode* |
|------|------------|---------|---------------|-----|-----------|
| 1**  | Exact      | Exact   | Exact         | Exact | Exact     |
| 2    | Soundex    | Soundex | Exact         | Exact | Exact     |
| 3    | 1st character | Characters 1–3 | Exact | Exact | Exact     |
| 4    | 1st character | Characters 1–3 | Exact | Exact | Exact     |
| 5    | Exact      | Exact   | Exact         | Exact | Exact     |
| 6    | Partial    | Exact   | Exact         | Exact | Exact     |
| 7    | Exact      | Exact   | Exact         | Exact | Exact     |
| 8    | 1st character | Characters 1–3 | Exact | Exact | Exact     |

Notes: * Full postcode (e.g. LS0 0AA). ** Step 1 was repeated by NHS Digital but allowing an NPD record to link to many PDS records. The objective of repeating this modified step 1 was to remove potential duplicate HESIDs for the same pupil. See details in Supplementary Appendix 4.

Exact refers to exact linking; Partial refers exact linking but using month and year of birth only; Soundex refers to the Structured Query Language (SQL) algorithm that converts an alphanumeric string to a four-character code that is based on how the string sounds when spoken. NPD = National Pupil Database; PDS = Personal Demographic Service.

Table 3: Linkage stage 2: 7 step deterministic algorithm for linking the personal demographic service to hospital episode statistics

| Step | NHS number | Date of birth | Sex | Postcode* |
|------|------------|---------------|-----|-----------|
| 1    | Exact      | Exact         | Exact | Exact     |
| 2    | Exact      | Exact         | Exact | Exact     |
| 3    | Exact      | Partial       | Exact | Exact     |
| 4    | Exact      | Partial       | Exact | Exact     |
| 5    | Exact      |               | Exact | Exact     |
| 6    | Exact      | Exact         | Exact | Exact     |

Where NHS number does not contradict the match and date of birth is not 1 January

| Step | NHS number | Date of birth | Sex | Postcode* |
|------|------------|---------------|-----|-----------|
| 7    | Exact      | Exact         | Exact | Exact     |

Where date of birth is not 1 January

Notes: * Full postcode (e.g. LS0 0AA). Exact refers to exact linking; Partial refers exact linking but using month and year of birth only.

to August) ordered from 2004/05 to 2016/17 for all pupils. In order to allow for multiple links with the highest level of agreement between NPD and PDS, step 1 was repeated (details in Supplementary Appendix 4). For all other steps, a pupil was removed from the linking pool (i.e. all records for that pupil were excluded from subsequent linking steps) once a linkage was identified.

Stage 2 involved linking the PDS table of identifiers for children linked to NPD with HES, using the NHS Digital internal 7 step algorithm (Table 3). The bridging files resulting from this linkage did not contain any identifiable data (such as name or date of birth) and contained all possible linkage pairs (linked and unlinked) resulting from linkage stages 1 and 2. Files contained the pseudonymised HESIDs for each of the four cohorts that included: all individuals in HES with a birth date in the relevant cohort and for those that linked to NPD, the anonymised PMR, two record-level indicators identifying the resulting linkage step of the linkage stages 1 and 2, and a variable indicating the specific cohort.

Figure 1 shows the transfer of pseudonymised HES attribute data (admitted patient care, accident and emergency, outpatient), together with the linkage bridging file of all possible linkage pairs, to the ONS SRS. Similarly, the Department for Education transferred NPD attribute data extracts containing the anonymised PMR to the ONS SRS.

The final phase of the process was to merge NPD and HES attribute data, using the bridging file obtained from stage 2 of the linkage. This was done by an Accredited Researcher (NL) in the ONS SRS. There were minor differences in HESIDs transferred by NHS Digital to UCL and those held by UCL as the NHS Digital HES data is continually updated, whereas UCL holds a static subset of the NHS Digital HES data (e.g. that is limited by age).

Evaluation of linkage quality

Among pupils who linked to a HES record, we calculated the distribution linked at each step for linkage stages 1 and 2, according to region, ethnic group, decile of deprivation, measured by income deprivation affecting children index (IDACI), and cohort year. We calculated the overall linkage rate as the percentage of pupils in the NPD who linked to any HES record for each of the four cohorts [27].

To evaluate potential bias resulting from missed matches, we compared characteristics of pupils in NPD who were linked to HES records with pupils in NPD who were not linked to HES [15, 28]. Unlinked pupils could include pupils who...
Figure 3: Results of linkage at stage 1 (NPD and PDS) and stage 2 (PDS and HES) and final linkage rates

**Notes:** NPD = national pupil dataset; PDS = personal demographics service; HES = hospital episode statistics; NHS = national health service; NHSD = NHS digital; ONS SRS = office for national statistics secure research service; UCL = university college London; aPMR = anonymised pupil matching reference.

never attended hospital or missed matches of pupils who did attend hospital. We used standardized differences (mean difference in standard deviation units) as these are thought to be more informative to detect potential biases than P-values in large samples [28, 29]. Standardized differences were calculated using the ‘stddiff’ command in Stata for the following variables: sex; ethnic group; region of pupil’s residence; IDACI Deciles; age at start of the first academic year; whether a child receives Special Education Need (SEN) provision (recorded in NPD as receiving Action, Action Plus or Support (AAP/S) and having a statement of SEN or an Education Health & Care Plan (S/EHCP) [30]); and persistent authorized annual absence rate for all academic years available defined as whether a child was absent in 10% or more of academic sessions (see Supplementary Appendix 5 for recording of variables) [31].

Multivariable logistic analysis was used to evaluate linkage from NPD to HES using the following demographic characteristics: sex, ethnicity, region of residence and IDACI Deciles.

**Results**

The bridging file produced by NHS Digital included 2,289,587 records with all possible linkage results. From this file, 41 duplicates were excluded since the same aPMR-HESID pairs linked in two different academic years. The second bridging file that included only the modified linkage step 1 of linkage stage 1 (i.e., where multiple links were allowed for each NPD record) contained 2,093,787 records, of which only 8,858 records were
Figure 4: Cumulative percentage of records linked in stage 1 (NPD to PDS; y axis) by academic year in spring census (x axis)

Notes: NPD = national pupil dataset; PDS = personal demographics service; HES = hospital episode statistics; NHS = national health service.

The registration online system (RON) is a web-based system registering life events (births and deaths) that was first piloted in November 2006 and fully implemented in July 2009. Since the implementation of RON, validation checks of addresses and postcodes have become possible at the point of registration [32]. Prior to the 2013/14 financial year, birth admissions were missing due to an extraction error by NHS Digital, resulting in postcodes missing in recorded birth episodes [33].

Distribution of pupil characteristics in linked records

At stage 1, between 91% and 95% of pupils linked at the first step of the 8-step algorithm, i.e. exact linkage by first name, surname, date of birth, sex and postcode (Table 2; Supplementary Appendix 6). However, evaluation by ethnic group showed that the additional steps in this algorithm, i.e. from 2–8, captured a greater percentage of ethnic minority groups (11.8% of minority ethnic groups versus 4.2% of white ethnic group).

A considerable percentage of records were linked in years after the first available Spring census (Figure 4). For example, 12% and 21% of records of pupils born in academic years 1990/91 and 1996/97 respectively, were matched after 2004/05 – their first available Spring census when it was possible to link to PDS. Similarly, in academic years 1999/00 and 2004/05, 16% and 9% of pupils were matched after their academic Year 1- their second available Spring census. For pupils born in academic year 1999/00 or after, the majority of records were linked in the first two academic years. In particular, 50% of records in cohort 1999/00 and 51% in 2004/05 were linked in Year 1, while 34% and 40% were linked in reception year (Supplementary Appendix 6).

Linkage rates by demographic characteristics of pupils

Pupils who linked to HES after both linkage stages and who were merged with HES attribute data comprise the matched dataset used for all subsequent analyses. Linkage rate by region, ethnic group, sex and IDACI deciles are shown in the Supplementary Appendix 7. We found that linkage rates...
Table 4: Sociodemographic characteristics of the pupil sample from the national pupil database linked and non-linked to hospital episode statistics (N = 2,294,369 pairs).

| Region                        | Cohort 1990/91 | Cohort 1996/97 | Stand. Diff. | Cohort 1999/00 | Cohort 2004/05 | Stand. Diff. |
|-------------------------------|----------------|----------------|--------------|----------------|----------------|--------------|
|                               | Non-linked     | Linked         |              | Non-linked     | Linked         |              |
|                               | (n = 47,934) (%) | (n = 565,798) (%) |              | (n = 35,299) (%) | (n = 536,619) (%) |              |
| Region                        | Non-linked     | Linked         | Stand.       | Non-linked     | Linked         | Stand.       |
|                               | (n = 22,185) (%) | (n = 507,725) (%) | Diff. (%)    | (n = 8,477) (%) | (n = 570,332) (%) | Diff. (%)    |
| Region                        | Non-linked     | Linked         | Stand.       | Non-linked     | Linked         | Stand.       |
|                               | (n = 47,934) (%) | (n = 565,798) (%) |              | (n = 35,299) (%) | (n = 536,619) (%) |              |
| London                        | 7,729 (16.1)   | 68,073 (12.0)  | 0.191        | 6,243 (17.7)   | 71,652 (13.4)  | 0.247        |
| South East                    | 8,000 (16.7)   | 81,806 (14.5)  |              | 5,961 (16.9)   | 75,452 (14.1)  |              |
| South West                    | 4,217 (8.8)    | 52,018 (9.2)   |              | 3,021 (8.6)    | 50,302 (9.4)   |              |
| West Midlands                 | 4,915 (10.3)   | 63,013 (11.1)  |              | 3,392 (9.6)    | 60,027 (11.2)  |              |
| North West                    | 6,200 (12.9)   | 83,376 (14.7)  |              | 3,630 (10.3)   | 77,805 (14.5)  |              |
| North East                    | 1,567 (3.3)    | 29,318 (5.2)   |              | 1,025 (2.9)    | 27,374 (5.1)   |              |
| Yorkshire and The Humber      | 3,885 (8.1)    | 57,539 (10.2)  |              | 2,908 (8.2)    | 54,564 (10.2)  |              |
| East Midlands                 | 3,535 (7.4)    | 47,096 (8.3)   |              | 2,769 (7.8)    | 42,187 (7.9)   |              |
| East of England               | 5,541 (11.6)   | 59,686 (10.5)  |              | 4,525 (12.8)   | 54,424 (10.1)  |              |
| Wales                         | 28 (0.1)       | 38 (0.0)       |              | 28 (0.1)       | 38 (0.0)       |              |
| Missing                       | 2,317 (4.8)    | 23,835 (4.2)   |              | 1,818 (5.2)    | 22,794 (4.2)   |              |
| Ethnic group                  |                |                |              |                |                |              |
| White                         | 27,692 (57.8)  | 488,330 (86.3) | 0.160        | 24,452 (69.3)  | 453,764 (84.6) | 0.159        |
| Asian                         | 2,541 (5.3)    | 33,024 (5.8)   |              | 2,584 (7.3)    | 37,654 (7)     |              |
| Black                         | 1,507 (3.1)    | 17,047 (3.0)   |              | 1,429 (4.0)    | 19,228 (3.6)   |              |
| Chinese                       | 278 (0.6)      | 1,384 (0.2)    |              | 213 (0.6)      | 1,439 (0.3)    |              |
| Other ethnic group            | 498 (1.0)      | 3,627 (0.6)    |              | 626 (1.8)      | 3,951 (0.7)    |              |
| Mixed                         | 834 (1.7)      | 13,808 (2.4)   |              | 1,278 (3.6)    | 19,286 (3.6)   |              |
| Missing                       | 14,584 (30.4)  | 8,578 (1.5)    |              | 4,717 (13.4)   | 1,297 (0.2)    |              |
| Sex                           |                |                |              |                |                |              |
| Male                          | 27,334 (57.0)  | 285,716 (50.5) | 0.131        | 17,014 (48.2)  | 275,479 (51.3) | 0.062        |
| Female                        | 20,543 (42.9)  | 279,520 (49.4) |              | 18,268 (51.8)  | 261,094 (48.7) |              |
| Missing                       | 57 (0.1)       | 562 (0.1)      |              | 17 (0.0)       | 46 (0.0)       |              |
| IDACI Deciles                 |                |                |              |                |                |              |
| 1 (deprived)                  | 7,306 (15.2)   | 54,336 (9.6)   | 0.242        | 4,866 (13.8)   | 50,540 (9.4)   | 0.218        |
| 2                             | 6,001 (12.5)   | 55,660 (9.8)   |              | 4,247 (12.0)   | 51,132 (9.5)   |              |
| 3                             | 5,414 (11.3)   | 56,149 (9.9)   |              | 3,811 (10.8)   | 51,662 (9.6)   |              |
| 4                             | 4,941 (10.3)   | 56,600 (10.0)  |              | 3,738 (10.6)   | 51,725 (9.6)   |              |
| 5                             | 4,611 (9.6)    | 56,620 (10.0)  |              | 3,444 (9.8)    | 52,336 (9.8)   |              |
| 6                             | 4,255 (8.9)    | 56,927 (10.1)  |              | 3,310 (9.4)    | 52,503 (9.8)   |              |
| 7                             | 3,854 (8.0)    | 56,891 (10.1)  |              | 2,936 (8.3)    | 53,336 (9.9)   |              |
| 8                             | 3,685 (7.7)    | 56,122 (9.9)   |              | 2,914 (8.3)    | 54,281 (10.1)  |              |
| 9                             | 3,514 (7.3)    | 54,878 (9.7)   |              | 2,851 (8.1)    | 55,791 (10.4)  |              |
| 10 (affluent)                 | 3,630 (7.6)    | 54,286 (9.6)   |              | 2,701 (7.7)    | 56,355 (10.5)  |              |
| Missing                       | 723 (1.5)      | 7,386 (1.3)    |              | 481 (1.4)      | 6,958 (1.3)    |              |
| Sex                           |                |                |              |                |                |              |
| Male                          | 27,334 (57.0)  | 285,716 (50.5) | 0.131        | 17,014 (48.2)  | 275,479 (51.3) | 0.062        |
| Female                        | 20,543 (42.9)  | 279,520 (49.4) |              | 18,268 (51.8)  | 261,094 (48.7) |              |
| Missing                       | 57 (0.1)       | 562 (0.1)      |              | 17 (0.0)       | 46 (0.0)       |              |

Libuy, N et al. International Journal of Population Data Science (2021) 6:1:13
Table 4: (Continued)

|                        | Cohort 1999/00 | Cohort 2004/05 |
|------------------------|---------------|----------------|
|                        | Non-linked    | Linked         | Stand. Diff. | Non-linked    | Linked         | Stand. Diff. |
| Wales                  | (n = 22,185)  | (n = 507,725)  |              | (n = 8,477)   | (n = 570,332)  |              |
| Missing                | 1,327 (6.0)   | 23,720 (4.7)   |              | 685 (8.1)     | 30,840 (5.4)   |              |
| Ethnic group           |               |               |              |               |               |              |
| White                  | 15,692 (70.7) | 415,660 (81.9) | 0.281        | 5,255 (62.0)  | 439,397 (77.0) | 0.358        |
| Asian                  | 2,581 (11.6)  | 43,061 (8.5)   |              | 1,207 (14.2)  | 57,790 (10.1)  |              |
| Black                  | 1,735 (7.8)   | 21,528 (4.2)   |              | 696 (8.2)     | 31,656 (5.6)   |              |
| Chinese                | 172 (0.8)     | 1,530 (0.3)    |              | 89 (1.0)      | 2,038 (0.4)    |              |
| Other ethnic group     | 700 (3.2)     | 5,146 (1.0)    |              | 486 (5.7)     | 8,375 (1.5)    |              |
| Mixed                  | 1,178 (5.3)   | 20,177 (4.0)   |              | 575 (6.8)     | 29,871 (5.2)   |              |
| Missing                | 127 (0.6)     | 623 (0.1)      |              | 169 (2.0)     | 1,205 (0.2)    |              |
| Sex                    |               |               |              |               |               |              |
| Male                   | 9,717 (43.8)  | 261,398 (51.5) | 0.153        | 3,660 (43.2)  | 292,784 (51.3) | 0.166        |
| Female                 | 12,445 (56.1) | 246,116 (48.5) |              | 4,814 (56.8)  | 277,508 (48.7) |              |
| Missing                | 23 (0.1)      | 211 (0.0)      |              | 0 (0.0)       | 43 (0.0)       |              |
| IDACI Deciles          |               |               |              |               |               |              |
| 1 (deprived)           | 2,863 (12.9)  | 49,733 (9.8)   | 0.142        | 909 (10.7)    | 53,590 (9.4)   | 0.07         |
| 2                      | 2,487 (11.2)  | 49,457 (9.7)   |              | 855 (10.1)    | 53,748 (9.4)   |              |
| 3                      | 2,257 (10.2)  | 49,130 (9.7)   |              | 849 (10.0)    | 54,246 (9.5)   |              |
| 4                      | 2,263 (10.2)  | 49,153 (9.7)   |              | 750 (8.8)     | 54,250 (9.5)   |              |
| 5                      | 2,139 (9.6)   | 49,450 (9.7)   |              | 812 (9.6)     | 55,571 (9.7)   |              |
| 6                      | 2,056 (9.3)   | 49,965 (9.8)   |              | 840 (9.9)     | 56,601 (9.9)   |              |
| 7                      | 1,980 (8.9)   | 50,467 (9.9)   |              | 844 (10.0)    | 57,776 (10.1)  |              |
| 8                      | 2,077 (9.4)   | 51,321 (10.1)  |              | 885 (10.4)    | 58,854 (10.3)  |              |
| 9                      | 1,972 (8.9)   | 52,884 (10.4)  |              | 858 (10.1)    | 61,048 (10.7)  |              |
| 10 (affluent)          | 1,953 (8.8)   | 53,904 (10.6)  |              | 821 (9.7)     | 62,514 (11.0)  |              |
| Missing                | 138 (0.6)     | 2,261 (0.4)    |              | 54 (0.6)      | 2,134 (0.4)    |              |

Notes: IDACI = Income deprivation affecting children index. Stand. Diff. = Standardized Difference.* Value omitted to avoid risk of disclosure due to small cell count.

improved over time for all these variables. However, ethnic minorities and pupils living in more deprived areas were less likely to match to HES. The linkage rate for white pupils improved from 94.6% in the 1990/91 cohort to 98.9% in the 2004/05 cohort. In contrast, for ethnic minority pupils in the same cohorts the linkage rate rose from 92.4% to 97.7%, respectively. We found a similar pattern by IDACI deciles. Linkage rates by region provide evidence that London has consistently lower linkage rates than the rest of the country.

Comparing characteristics of linked and unlinked pupils

Differences in the distribution of sociodemographic and educational characteristics of pupils recorded in NPD who linked or not to HES are shown in Table 4 (and Supplementary Table 9.1–9.4 in Supplementary Appendix 8). Overall, relatively low standardized differences are observed across all variables providing evidence of small or moderate differences between linked and unlinked groups. We considered standardized differences of 0.2, 0.5 and 0.8 as small, moderate and large, respectively [28, 34]. The largest differences were for the AAP/S and persistent authorized absence rate in cohort 1996/97 with values of 0.44 and 0.42. The mean standardized difference across cohort for region and ethnic groups was 0.25 and 0.24 whereas for sex and IDACI deciles was 0.13 and 0.17 (Table 4).

Evaluation of linkage from NPD to HES

Table 5 shows the results of multivariable logistic models displaying adjusted Odds Ratios (OR) for linkage to HES. Unadjusted models are also shown in Supplementary Appendix 9. OR below 1 indicates lower odds of linkage to HES compared with the reference category. Consistent with linkage rate estimates, we found differences across ethnic groups, deprivation and region. Across all cohorts, we found that relative to pupils of white ethnicity, pupils of ethnic minorities including Asian, Black, Chinese, Mixed and Any other ethnic group were less like to be matched. The odds of linkage to HES for Asian ethnic groups were less than ethnic minority pupils (e.g. 1990/91: Adjusted OR 0.69, 95% CI 0.66 to 0.72, p < 0.01; 2004/05: Adjusted OR 0.51, 95% CI 0.47 to 0.54, p < 0.01). Relative to male pupils, with the exception of pupils born in academic year 1990/91, female pupils were less likely to be matched (e.g. 2004/05: Adjusted OR 0.72, 95% CI 0.69 to 0.75, p < 0.01). Compared to pupils in the fifth IDACI Deciles,
Table 5: Adjusted odds ratios for a link between NPD and HES records according to sociodemographic characteristics in the NPD

| Characteristics from NPD | Cohort 1990/91 | Cohort 1996/97 |
|--------------------------|----------------|----------------|
|                          | aOR Conf. Int. | aOR Conf. Int. |
| **Ethnic group**         |                |                |
| White                    | Ref            | Ref            |
| Asian                    | 0.69 [0.66,0.72]** | 0.69 [0.66,0.73]** |
| Black                    | 0.62 [0.59,0.66]** | 0.67 [0.63,0.71]** |
| Chinese                  | 0.29 [0.26,0.33]** | 0.38 [0.33,0.44]** |
| Any other ethnic group   | 0.42 [0.38,0.46]** | 0.32 [0.30,0.35]** |
| Mixed                    | 0.92 [0.85,0.98]** | 0.80 [0.75,0.85]** |
| Missing                  | 0.03 [0.03,0.03]** | 0.01 [0.01,0.02]** |
| **Sex**                  |                |                |
| Male                     | Ref            | Ref            |
| Female                   | 1.35 [1.32,1.37]** | 0.87 [0.85,0.89]** |
| Missing                  | 22.77 [17.02,30.47]** | 10.21 [5.77,18.07]** |
| **Region**               |                |                |
| London                   | Ref            | Ref            |
| South East               | 1.31 [1.26,1.36]** | 1.12 [1.08,1.17]** |
| South West               | 1.34 [1.28,1.40]** | 1.38 [1.31,1.45]** |
| West Midlands            | 1.27 [1.22,1.33]** | 1.37 [1.30,1.43]** |
| North West               | 1.36 [1.30,1.41]** | 1.64 [1.57,1.72]** |
| North East               | 1.91 [1.80,2.04]** | 1.99 [1.85,2.14]** |
| Yorkshire and The Humber | 1.34 [1.28,1.40]** | 1.42 [1.35,1.49]** |
| East Midlands            | 1.28 [1.23,1.35]** | 1.22 [1.16,1.28]** |
| East of England          | 1.14 [1.09,1.19]** | 1.00 [0.95,1.04]** |
| Wales                    | 0.31 [0.16,0.59]** | 0.40 [0.17,0.93]** |
| Missing                  | 1.16 [1.10,1.23]** | 1.08 [1.01,1.14]** |
| IDACI Deciles            |                |                |
| 1 (deprived)             | 0.67 [0.64,0.70]** | 0.71 [0.67,0.74]** |
| 2                        | 0.78 [0.74,0.81]** | 0.77 [0.73,0.81]** |
| 3                        | 0.86 [0.82,0.90]** | 0.87 [0.83,0.92]** |
| 4                        | 0.95 [0.90,0.99]** | 0.90 [0.85,0.94]** |
| 5                        | Ref            | Ref            |
| 6                        | 1.11 [1.05,1.16]** | 1.05 [1.00,1.11]** |
| 7                        | 1.26 [1.20,1.32]** | 1.23 [1.16,1.29]** |
| 8                        | 1.31 [1.25,1.38]** | 1.27 [1.20,1.34]** |
| 9                        | 1.31 [1.25,1.38]** | 1.37 [1.29,1.44]** |
| 10 (affluent)            | 1.27 [1.21,1.34]** | 1.52 [1.44,1.61]** |
| Missing                  | 0.95 [0.86,1.04] | 1.06 [0.95,1.18] |
| Observations             | 613,732         | 571,918        |
| Pseudo R-squared         | 0.162           | 0.093          |

| Characteristics from NPD | Cohort 1999/00 | Cohort 2004/05 |
|--------------------------|----------------|----------------|
|                          | aOR Conf. Int. | aOR Conf. Int. |
| **Ethnic group**         |                |                |
| White                    | Ref            | Ref            |
| Asian                    | 0.56 [0.54,0.59]** | 0.51 [0.47,0.54]** |
| Black                    | 0.43 [0.40,0.45]** | 0.47 [0.43,0.51]** |
| Chinese                  | 0.35 [0.30,0.41]** | 0.27 [0.22,0.34]** |
| Any other ethnic group   | 0.26 [0.24,0.28]** | 0.18 [0.17,0.20]** |
| Mixed                    | 0.64 [0.60,0.68]** | 0.60 [0.55,0.66]** |
| Missing                  | 0.21 [0.17,0.25]** | 0.09 [0.07,0.10]** |
| **Sex**                  |                |                |
| Male                     | Ref            | Ref            |
| Female                   | 0.73 [0.71,0.75]** | 0.72 [0.69,0.75]** |
| Missing                  | 0.61 [0.39,0.96]** | 1.00 [0.29,3.50]** |
Libuy, N et al. International Journal of Population Data Science (2021) 6:1:13

Table 5: (Continued)

| Characteristics from NPD | Cohort 1999/00 | Cohort 2004/05 |
|-------------------------|--------------|--------------|
|                         | aOR          | Conf. Int.   | aOR          | Conf. Int.   |
| Region                  |              |              |              |              |
| London                  | Ref          | Ref          |              |              |
| South East              | 1.00         | [0.95,1.04]  | 0.94         | [0.87,1.02]  |
| South West              | 1.62         | [1.51,1.73]**| 1.38         | [1.24,1.54]**|
| West Midlands           | 1.23         | [1.16,1.30]**| 1.21         | [1.11,1.33]**|
| North West              | 1.64         | [1.55,1.74]**| 1.09         | [1.00,1.19]* |
| North East              | 1.82         | [1.67,2.00]**| 1.71         | [1.47,1.99]**|
| Yorkshire and The Humber| 1.61         | [1.51,1.72]**| 1.23         | [1.12,1.35]**|
| East Midlands           | 1.14         | [1.08,1.21]**| 0.96         | [0.87,1.06]  |
| East of England         | 0.86         | [0.81,0.90]**| 0.83         | [0.76,0.90]**|
| Wales                   | 0.37         | [0.17,0.80]**| 0.36         | [0.11,1.19]  |
| Missing                 | 0.97         | [0.91,1.03]  | 0.74         | [0.68,0.82]**|

| IDACI Deciles           |              |              |              |              |
| 1 (deprived)            | 0.73         | [0.68,0.77]**| 0.82         | [0.75,0.90]**|
| 2                       | 0.82         | [0.77,0.87]**| 0.88         | [0.80,0.97]* |
| 3                       | 0.89         | [0.83,0.94]**| 0.90         | [0.82,1.00]**|
| 4                       | 0.92         | [0.86,0.97]**| 1.03         | [0.93,1.14]  |
| 5                       | Ref          |              | Ref          |              |
| 6                       | 1.10         | [1.03,1.17]**| 1.03         | [0.94,1.14]  |
| 7                       | 1.18         | [1.11,1.26]**| 1.11         | [1.00,1.22]**|
| 8                       | 1.20         | [1.13,1.28]**| 1.14         | [1.03,1.25]**|
| 9                       | 1.36         | [1.27,1.45]**| 1.31         | [1.18,1.45]**|
| 10 (affluent)           | 1.48         | [1.39,1.58]**| 1.57         | [1.42,1.74]**|
| Missing                 | 0.87         | [0.73,1.05]  | 0.80         | [0.59,1.07]  |
| Observations            | 529,910      | 578,809      |
| Pseudo R-squared        | 0.026        | 0.027        |

Notes: Adjusted for all other covariates listed in the table. *p < 0.05, **p < 0.01. aOR = adjusted odds ratios. Conf. Int. = confidence interval. NPD = national pupil dataset. HES = hospital episode statistics; NHS = national health service. IDACI = income deprivation affecting children index.

pupils living in the most deprived areas were less likely to be matched, whereas pupils living in the most affluent areas were more likely to be matched. Similarly, results for the region of pupil residence show differences for linkage success.

Discussion

This study is the first to link administrative records from schools and hospitals for all children and adolescents attending state-funded schools in England for four 1-year birth cohorts (~2.2 million children). It builds upon previous studies that have demonstrated the public benefit and challenges for data sharing across educational and health services for specific subgroups [8, 13, 35, 36], and in other countries [9–14]. We evaluated two deterministic algorithms implemented by NHS Digital and found that although linkage rates were high and improved over time, pupils from ethnic minority groups or living in areas of high deprivation were disproportionately less likely to match to HES.

Key findings

Our finding that the linkage rate was 99% for the youngest cohort is encouraging for future studies using multi-step deterministic algorithms in England. This linkage rate is similar to studies in Scotland, Wales and Australia that used probabilistic linkage methods [11, 13, 14, 37–39]. For instance, linkage rates for the annual Scottish Governments pupils census linked to the community health index database ranged between 86.3% and 95% [14], while two other Scottish studies found linkage rates of 99.7% [13] and 81.8% [11].

We found that between 2.3–7.6% of ethnic minority pupils were not linked to health records. Ethnic differences reported in previous linkage success reflect differences in the quality of registration of Chinese, Asian and Hispanic names [8, 27, 28]. The differences in linkage rates by ethnic minority in linkage steps that relaxed the requirement to agree on exact full name suggest that inconsistencies in forenames and surnames explain the lower linkage rates for ethnic minority pupils. Residential instability may also be relevant: lower rates of linkage for pupils from ethnic minorities at steps 1 and 2 between PDS and HES (i.e. stage 2), could be due to poor recording of postcode, as reported in other studies [40, 41]. It is also estimated that 20% of children aged 0 to 15 years are born outside the UK, which may have a differential impact on linkage success [42]. Additional steps in the deterministic algorithm that incorporate phonetic systems codes for other languages [43, 44], or methods that discriminate partial agreements in string comparisons [45–48], or probabilistic
linkage methods could be used to further improve linkage rates for ethnic minorities [40, 48].

We found that pupils living in more deprived neighbourhoods were less likely to link to health records than pupils living in more affluent areas. Previous studies have suggested that families from more affluent areas are more likely to comply with the administrative process [8]. However, pupils living in London were less likely to link to HES records than in other regions, even after accounting for sociodemographic characteristics. This difference may reflect higher rates of international emigration from London, less use of health services, differential use of private health services, or poorer quality of identifiers in London.

Improvements in the quality of recording of identifiers in schools and health data systems likely account for improved linkage rates over time. Changes in health systems governing collection of patient identifiers, such as the implementation of NHS Numbers for Babies (NN4B) service on 29th October 2002, the introduction of Registration ONline system (RON) on 1st July 2009, the correction of a postcode extraction error by NHS Digital on 1st April 2013, have been shown to improve the completeness of identifiers used in the linkage [20]. Retrospective correction of this extraction error and re-linkage by NHS Digital of birth episodes to subsequent HES records, would be expected to improve linkage to NPD in earlier years.

Strengths and limitations

Our study demonstrates very high linkage rates between educational and HES records for pupils attending state schools in England. The governance for this project addressed the challenges of cross-sectoral linkage between health and educational institutions in England whilst avoiding disclosure during the linkage process [16]. Use of multiple steps at each stage of linkage, and of identifiers recorded over multiple years for each child, were critical to achieving high linkage rates. Preliminary findings indicate that two-thirds of the linked HES records related to at least one admission, excluding the birth admission (to be reported elsewhere). The linkage algorithms used for this project are currently being used to link educational and health records for all pupils in England born academic years 1995/96 onwards and will be relevant for other studies linking data to HES or NPD (or both) [17].

Linking educational data with hospital and death records creates new possibilities for studying a wide spectrum of policy-relevant questions. For example, the availability of data across the child life course could enable studies into the impact of health on education and education on health. Linked data for all children will be made available for applications for research from government and academia in 2021 [49, 50].

Record-level indicators of the linkage process (i.e. variables indicating the step in our rule-based linkage algorithms at which a pair of records linked) were shared by NHS Digital to enable us to evaluate linkage biases. We used this information to demonstrate the value of later steps in the algorithm for linking pupils from ethnic minority and deprived areas. However, we did not have information on country of birth, and so could not assess whether linkage rates were lower for children who were born outside England. Future studies should consider sharing information about the completeness or quality of the identifiers to identify whether changes in data entry systems could address missed links in these more vulnerable groups [16].

A limitation and advantage were the system changes in administrative data resulting in improvements in identifier and linkage quality and additional data collections from both services. These changes can introduce variation in linkage error over time, for instance, patients with fewer contacts with health services or more mobile populations could have out-of-date residential information in PDS disproportionately affecting linkage quality, which analysts need to consider when investigating trends.

A further limitation is that since no gold-standard dataset defining true match status was available, we could not derive standard measures of linkage quality (sensitivity/recall, false match rate and positive predictive value/precision). Approaches for estimating rates of false matches in further linkage between HES and NPD could be applied, for example by applying the linkage algorithms to a set of ‘negative controls’ (i.e. NPD records for which we are certain there should be no link in HES or vice versa) and counting how many records were erroneously linked [51, 52]. This would allow an estimation of false match rates, but would not allow identification of which records were falsely matched. Existing ‘gold-standard’ data for health records in England for specific sub populations also have the potential to be used in the future evaluations of linkage quality [53]. Future studies could develop representative gold-standard data using known links from UK cohort studies, such as the Millennium Cohort Study or Next Steps to allow linkage error to be fully measured [54, 55].

Implications

We created a de-identified linked database that brings together data from the Department for Education (education and social care) and hospitalisation data for all children – the ECHILD Database. This resource will be made available for approved researchers later in 2021 for purposes that benefit health, wellbeing, education and the provision of health or social care. The ECHILD dataset will enable a step change in the scale and depth of research into the inter-relationships between health, education and social care across the life course, and how services across England vary in their responses.

Our linkage created a de-identified bridging file that combines pseudo-identifiers from education (anonymised Pupil Matching Reference) and HES. This bridging file can be used by the data providers to link to further datasets for approved studies, without the need to link real-world identifiers such as names and postcodes. As the data systems for capturing identifiers change, as is currently happening at NHS Digital [56], our evaluation of linkage success will need to be repeated and linkage metrics provided to researchers.

Researchers addressing questions relating to ethnic minority or deprived groups need to consider whether to adjust for missing data among these groups due to missed links. Statistical techniques include weighting or imputation, depending on the research objectives [57].
Conclusion

We found high linkage rates between administrative education and hospital data for pupils in four cohorts born between academic years 1990/91-2004/05 in England. Linkage rates improved over time, but ethnic minorities and pupils living in deprived neighbourhoods were disproportionally affected by linkage error. Evidence from comparing linked and unlinked populations provides measures that can be used to take into account potential biases due to linkage error.

Acknowledgements

This research benefits from and contributes to the NIHR Children and Families Policy Research Unit, but was not commissioned by the National Institute for Health Research (NIHR) Policy Research Programme. We are grateful to Gary Connell (Department for Education), Garry Coleman (NHS Digital) and their teams for supporting this work. We thank the ECHILD team: Dr. David Etoori, Dr. Louise Mc Grath-Lone, Matthew Lilliman and Dr Erin Walker.

Funding

This work was supported by ESRC via the Administrative Data Research UK through the Strategic Hub [grant number ES/V000977/1]; the Administrative Data Research Centre for England; the NIHR Great Ormond Street Hospital Biomedical Research Centre and the Health Data Research UK [grant number LOND1], which is funded by the UK Medical Research Council and eight other funders; Wellcome Trust [grant number 212953/Z/18/Z to KH]; and UKRI Innovation Fellowship funded by the Medical Research Council [grant number MR/S003797/1 to RB].

Data availability

The data underlying this article cannot be shared publicly due to data sharing agreements with NHS Digital and Department for Education.

Conflict of interest statement

None declared.

Ethics statement

Research ethics approval was granted (project ID 232547, REC reference 17/LO/1494) and data sharing agreements are in place with NHS Digital (NIC-27404) and the Department for Education (DR150701.02). The Confidentiality Advisory Group confirmed that this research is exempt from review (reference 15/CAG/0004) because it only uses pseudonymised NHS data.

Supplementary appendices

Supplementary Appendix 1: Description of Linkage bridging files transferred to UCL Data Safe Haven and to the Office of National Statistics Secure Research Service

Supplementary Appendix 2: Timelines of the four cohorts alongside availability of data from Hospital Episode Statistics, National Pupil Dataset data and Personal Demographics Service

Supplementary Appendix 3: Description of data resources used in the linkage.

Supplementary Appendix 4: Description of linkage between Personal Demographics Service and National Pupil Dataset

Supplementary Appendix 5: Description of demographic variables in National Pupil Dataset

Supplementary Appendix 6: Performance of linkage stages

Supplementary Appendix 7: Linking rates

Supplementary Appendix 8: Standardized differences and P-values

Supplementary Appendix 9: Linkage evaluation Logit models

References

1. 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(4):1093-i. https://doi.org/10.1093/ije/dyx015.

2. Jay MA, Grath-Lone LM, Gilbert R. Data Resource: the National Pupil Database (NPD). International Journal of Population Data Science. 2019;4(1). https://doi.org/10.23889/ijpds.v4i1.1101.

3. Crawford C, Dearden L, Greaves E. The drivers of month-of-birth differences in children’s cognitive and non-cognitive skills. J R Stat Soc Ser A Stat Soc. 2014;177(4):829–60. https://doi.org/10.1111/jrssa.12071.

4. Zylbersztejn A, Gilbert R, Hjern A, Wijlaars L, Hardelid P. Child mortality in England compared with Sweden: a birth cohort study. Lancet. 2018;391(10134):2008-18. https://doi.org/10.1016/S0140-6736(18)30670-6.

5. Herbert A, Gilbert R, Gonzalez-Izquierdo A, Li L. Violence, self-harm and drug or alcohol misuse in adolescents admitted to hospitals in England for injury: a retrospective cohort study. BMJ Open. 2015;5(2):e006079. https://doi.org/10.1136/bmjopen-2014-006079.

6. Coathup V, Boyle E, Carson C, Johnson S, Kurinzczuk JJ, Macfarlane A, et al. Gestational age and hospital admissions during childhood: population based, record linkage study in England (TIGAR study). BMJ. 2020;371:m4075. https://doi.org/10.1136/bmj.m4075.

7. Harron K, Gilbert R, Fagg J, Guttmann A, van der Meulen J. Associations between pre-pregnancy psychosocial risk factors and infant outcomes: a population-based cohort study in England. Lancet Public Health. 2021;6(2):e97–e105. https://doi.org/10.1016/S2468-2667(20)30210-3.
8. Downs JM, Ford T, Stewart R, Epstein S, Shetty H, Little R, et al. An approach to linking education, social care and electronic health records for children and young people in South London: a linkage study of child and adolescent mental health service data. BMJ Open. 2019;9(1):e024355. https://doi.org/10.1136/bmjopen-2018-024355.

9. Jones KH, Ford DV, Thompson S. A Profile of the SAIL Databank on the UK Secure Research Platform. International journal of population data science. 2019;4(2). https://doi.org/10.23889/ijpds.v4i2.1134.

10. Lynch J. The South Australian Early Childhood Data Project. School of Public Health: University of Adelaide; 2016 2016. Report No.: 4.

11. MacKay DF, Smith GCS, Dobbie R, Pell JP. Gestational Age at Delivery and Special Educational Need: Retrospective Cohort Study of 407,503 Schoolchildren. PLOS Medicine. 2010;7(6):e1000289. https://doi.org/10.1371/journal.pmed.1000289.

12. Maret-Ouda J, Tao W, Wahlin K, Lagergren J. Nordic registry-based cohort studies: Possibilities and pitfalls when combining Nordic registry data. Scand J Public Health. 2017;45(17_suppl):14–9. https://doi.org/10.1177/1403494817702336.

13. Stewart CH, Dundas R, Leyland AH. The Scottish school leavers cohort: linkage of education data to routinely collected records for mortality, hospital discharge and offspring birth characteristics. BMJ Open. 2017;7(7). https://doi.org/10.1136/bmjopen-2016-015027.

14. Wood R, Clark D, King A, Mackay D, Pell J. Novel cross-sectoral linkage of routine health and education data at an all-Scotland level: a feasibility study. The Lancet. 2013;382:S10. https://doi.org/10.1016/S0140-6736(13)62435-6.

15. Doidge JC, Harron KL. Reflections on modern methods: linkage error bias. Int J Epidemiol. 2019;48(6):2050–60. https://doi.org/10.1093/ije/dyz203.

16. Gilbert R, Lafferty R, Hagger-Johnson G, Harron K, Zhang L-C, Smith P, et al. GUILD: GUIdance for Information about Linking Data sets. J Public Health (Oxf). 2018;40(1):191–8. https://doi.org/10.1093/jpubhealth/pfx203.

17. ECHILD. The Education and Child Health Insights from Linked Data 2021 [Available from: https://www.ucl.ac.uk/child-health/research/population-policy-and-practice-research-and-teaching-department/cenb-clinical-20].

18. Green F, Anders J, Henderson M, Henseke G. Who Chooses Private Schooling in Britain and Why? In: Societies LCfLaLCiKEa, editor. Institute of education, UCL2017.

19. Harron K, Gilbert R, Cromwell D, van de Meulen J. Linking Data for Mothers and Babies in DeIdentified Electronic Health Data. PLoS One. 2016;11(10):e0164667. https://doi.org/10.1371/journal.pone.0164667.

20. Zylbersztejn AG, Ruth; Hardelid, Pia. Impact of changes to data collection on a national birth cohort from administrative health records in England. PLoS One. 2020.

21. NHS Digital. A Guide to Linked Mortality Data from Hospital Episode Statistics and the Office for National Statistics. 2015 [Available from: https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/linked-hes-ons-mortality-data#ons-mortality-datafrom:].

22. (HSCIC) THaSCIC. Methodology for creation of the HES Patient ID (HESID) 2014 [Available from: https://webarchive.nationalarchives.gov.uk/20180328130852tf_/http://content.digital.nhs.uk/media/1370/HES-Hospital-Episode-Statistics-Replacement-of-the-HES-patient-ID/pdf/HESID_Methodology.pdf/].

23. Commons ThO. Department of Health: The National Programme for IT in the NHS, Twentieth Report of Session 2006–07. The Stationery Office Limited: House of Commons, Committee of Public Accounts; 2007 11/04/2007.

24. Statistics O.N. Personal Demographics Service data Office for National Statistics2020 [Available from: https://www.ons.gov.uk/census/censustransformation programme/administrativedatacensusproject/datasource overviews/personaldemographicsservicedata].

25. Boyd A, Thomas R, Macleod J. NHS Number and the systems used to manage them: an overview for research users. Cohort & Longitudinal Studies Enhancement Resources (CLOSER): Population Health Sciences, Bristol Medical School, University of Bristol; 2018.

26. NHS Digital. Personal Demographics Service fair processing 2020 [Available from: https://digital.nhs.uk/services/demographics/personal-demographics-service-fair-processing#:text=The%20Personal%20Demographics%20Service%20(PDS,%20known%20as%20demographic%20information)].

27. Bohensky M. Bias in data linkage studies. In: Harron KG, Harvey; Dibben, Chris, editor. Methodological Developments in Data Linkage: Wiley; 2016. p. 63–82.

28. Harron KL, Doidge JC, Knight HE, Gilbert RE, Goldstein H, Cromwell DA, et al. A guide to evaluating linkage quality for the analysis of linked data. Int J Epidemiol. 2017;46(5):1699–710. https://doi.org/10.1093/ije/dyx177.

29. Austin PC. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. Stat Med. 2009;28(25):3083–107. https://doi.org/10.1002/sim.3697.

30. Jay MA, Gilbert R. Special educational needs, social care and health. Arch Dis Child. 2020. https://doi.org/10.1136/archdischild-2019-317985.
31. Bayoumi AM. STDDIFF: Stata module to compute standardized differences for continuous and categorical variables. In: Economics BCDo, editor. Statistical Software Components S4582752016.

32. Statistics OfN. Office for National Statistics. Mortality Statistics?: Metadata. Office for National Statistics.: The Office for National Statistics (ONS); 2015. p. 35–41.

33. 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(12):e0243843. https://doi.org/10.1371/journal.pone.0243843.

34. Austin PC. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. Stat Med. 2009;28(25):3083–107. https://doi.org/10.1002/sim.3697.

35. Fleming M, Fitton CA, Steiner MFC, McLay JS, Clark D, King A, et al. Educational and health outcomes of children and adolescents receiving antidepressant medication: Scotland-wide retrospective record linkage cohort study of 766,237 schoolchildren. Int J Epidemiol. 2020;49(4):1380–91. https://doi.org/10.1093/ije/dya002.

36. Fleming M, Fitton CA, Steiner MFC, McLay JS, Clark D, King A, et al. Educational and health outcomes of children treated for asthma: Scotland-wide record linkage study of 683,716 children. European Respiratory Journal. 2019;54(3). https://doi.org/10.1183/13993003.02309-2018.

37. Holman CDAJ, Bass AJ, Rouse IL, Hobbs MST. Population-based linkage of health records in Western Australia: development of a health services research linked database. Australian and New Zealand Journal of Public Health. 1999;23(5):453–9. https://doi.org/10.1111/j.1467-842X.1999.tb01297.x.

38. Jones KH, Ford DV, Jones C, Dsilva R, Thompson S, Brooks CJ, et al. A case study of the Secure Anonymous Information Linkage (SAIL) Gateway: A privacy-protecting remote access system for health-related research and evaluation. Journal of Biomedical Informatics. 2014;50:196–204. https://doi.org/10.1016/j.jbi.2014.01.003.

39. Wellcome Trust. Public health research data forum. enabling data linkage to maximise the value of public health research data: full report. 2015 [Available from: https://wellcome.ac.uk/sites/default/files/enabling-data-linkage-to-maximise-value-of-public-health-research-data-phrdf-mar15.pdf.]

40. Hagger-Johnson G, Harron K, Goldstein H, Aldridge R, Gilbert R. Probabilistic linkage to enhance deterministic algorithms and reduce data linkage errors in hospital administrative data. J Innov Health Inform. 2017;24(2):891. https://doi.org/10.14236/jhi.v24i2.891.

41. Roberts E, Doidge JC, Harron KL, Hotopf M, Knight J, White M, et al. National administrative record linkage between specialist community drug and alcohol treatment data (the National Drug Treatment Monitoring System (NDTMS)) and inpatient hospitalisation data (Hospital Episode Statistics (HES)) in England: design, method and evaluation. BMJ Open. 2020;10(11):e043540. https://doi.org/10.1136/bmjopen-2020-043540.

42. Krausove A, Vargas-Silva C. BRIEFING England: Censu Profile. The Migration Observatory, University of Oxford; 2014.

43. Zahoransky D, Polášek I. Text Search of Surnames in Some Slavic and Other Morphologically Rich Languages Using Rule Based Phonetic Algorithms. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2015;23(3):553–63. https://doi.org/10.1109/TASLP.2015.2393393.

44. Harron K. An Introduction to Data Linkage. In: ADRN Publication UoE, editor. Better Knowledge Better Society. The Administrative Data Research Network 2016.

45. Gong J, Wang L, Oard D. Matching person names through name transformation. Proceeding of the 18th ACM Conference on Information and Knowledge Management. 2009:1875–8.

46. Newcombe HB, Fair ME, Lalonde P. Discriminating powers of partial agreements of names for linking personal records. Part II: The empirical test. Methods Inf Med. 1989;28(2):92–6.

47. Newcombe HB, Fair ME, Lalonde P. Discriminating powers of partial agreements of names for linking personal records. Part I: The logical basis. Methods Inf Med. 1989;28(2):86–91.

48. Treeratpituk P, Giles CL. Name-Ethnicity Classification and Ethnicity-Sensitive Name Matching. Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence. 2012.

49. UK A. ECHILD: Linking children’s health and education data for England 2021 [Available from: https://www.adruk.org/our-work/browse-all-projects/echild-linking-childrens-health-and-education-data-for-england-142/.

50. London UC. Education and Child Health Insights from Linked Data 2021 [Available from: https://www.ucl.ac.uk/child-health/research/population-policy-and-practice-research-and-teaching-department/cenb-clinical-20.

51. Harron K, Doidge JC, Goldstein H. Assessing data linkage quality in cohort studies. Annals of Human Biology. 2020;47(2):218–26. https://doi.org/10.1080/03014460.2020.1742379.

52. Doidge J, Christen P, Harron K. Quality Review: Quality Assessment in Data Linkage. Government Analysis Function & Office for National Statistics; 2020.
53. Harron K, Wade A, Gilbert R, Muller-Pebody B, Goldstein H. Evaluating bias due to data linkage error in electronic healthcare records. BMC Medical Research Methodology volume. 2014;14(36). https://doi.org/10.1186/1471-2288-14-36.

54. University College London UioE, Centre for Longitudinal Studies, NHS Digital. . Next Steps: Linked Health Administrative Datasets (Hospital Episode Statistics), England, 1997–2017. UK Data Service (2020).

55. Hockley C, Quigley M, Hughes G, Calderwood L, Joshi H, Davidson L. Linking Millennium Cohort data to birth registration and hospital episode records. Paediatric and perinatal epidemiology. 2008;22:99–109. https://doi.org/10.1111/j.1365-3016.2007.00902.x.

56. NHS Digital. Hospital Episode Statistics data changes in 2021 2021 [Available from: https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/hospital-episode-statistics/hospital-episode-statistics-data-changes-in-2021.

57. Goldstein H, Harron K, Wade A. The analysis of record-linked data using multiple imputation with data value priors. Stat Med. 2012;31(28):3481–93. https://doi.org/10.1002/sim.5508.
Online Only Supplemental Appendix

Supplementary Appendix 1: Description of linkage bridging files transferred to UCL Data Safe Haven and to the Office of National Statistics Secure Research Service

As described in Figure 1, NHS Digital transferred two linkage bridging files to UCL Data Save Haven (DSH). The first file contained the results of the first two linkage stages (See Tables 1 and Tables 2). This file included the link results when steps in Table 1 allow to link an NPD record to one PDS record only.

To remove potential duplicate HES IDs for the same pupil, a second linkage bridging file was created by NHS Digital and sent to UCL DSH. This second file included the link result in which only a modified step 1 of Table 1 was run, allowing an NPD record to link to many PDS records. Steps 2-8 in Table 1 were not run, and steps 1–7 in Table 2 were repeated.

Supplementary Appendix 2: Timelines of the four cohorts alongside availability of data from Hospital Episode Statistics, National Pupil Dataset data and Personal Demographics Service

Supplementary Figure 1 shows timelines of the four cohorts alongside availability of data from HES (Grey), DfE NPD data (Green) and PDS (Yellow).

Cohort 1

This cohort includes young people born between 01/09/1990-31/08/1991 who entered reception class in September 1996. These young people are captured in HES on their first hospitalisation on or after 01/04/1997 (at approximately 7 years of age or more). These children are first recorded in the NPD with their KS1 and KS2 data from 1998 and 2002, respectively, and annual school census from 2001/2 onwards. This cohort tests linkage with all children (state and non-state educated) who have a KS4 assessment (approximately 99% of adolescents aged 15/16). The other cohorts only capture children attending state schools (around 92% of the population). The cohort also tests linkage with young people receiving higher education up to age 18 years, a group who are likely to move and who have relatively high rates of admission to hospital.

Cohort 2

This cohort comprises children born between 01/09/1996-31/08/1997 who entered reception class in September 2002. These children are recorded in the NPD annual school census from 2001/2 and in KS1 data from 2003/4. These children enter secondary school in September 2008 and have KS3 recorded in 2011 and KS4 recorded in 2012/13. Data on hospitalisations is captured in HES on or after 01/04/1997 (at least approximately 1 year of age) until the most recent data extract available. This cohort has annual Pupil Level Annual School Census (PLASC) census data throughout the primary school years, but not all have had a hospital admission.

Cohort 3

These children are born between 01/09/1999-31/08/2000 and enter NPD in the school census at reception in September 2005. This cohort captures indicators of chronic conditions recorded in the birth record and in infancy, which is the period when the risk of admission to hospital is the highest. These children are recorded in KS1, KS2, KS3 and KS4 data from 2006/07, 2010/11, 2013/14, and 2015/16, respectively.

Cohort 4

These children were born between 01/09/2004-31/08/2005 and entered NPD in the school census at reception in September 2010. NPD are requested up until KS2 data which will end in 2015/16. This cohort data in HES and PDS is concurrent, and therefore we expect a higher likelihood of successful linkage.
Supplementary Figure S1: Timelines of the four cohorts alongside availability of data from HES (Grey), DfE NPD data (Green) and PDS (Yellow)

Notes: This figure show the timelines of the four cohorts alongside availability of data from the three sources HES, PDS and NPD for each year on the 1 April. Cohort 1: young people born between 01/09/1990-31/08/1991 who entered reception class in September 1996. DfE NPD extracts include all pupils in Year 6 in 2001/02 as a proxy for those in reception year in 1995/96. Cohort 2: children born between 01/09/1996-31/08/1997 who entered reception class in September 2002. Cohort 3: children are born between 01/09/1999-31/08/2000 and enter NPD in the school census at reception in September 2005. Cohort 4: children are born between 01/09/2004-31/08/2005 and enter NPD in the school census at reception in September 2010. Numbers in last row for each cohort correspond to age of child. Black stars are point of entry to reception.

NPD National Pupil Dataset; PDS Personal Demographics Service; HES Hospital Episode Statistics; NHS National Health Service; ONS Office for National Statistics; APC Admitted Patient Care; A&E Accident and Emergency; EYC Early Year Census; EYFSP Early years foundation stage profile; KS Key Stage.
Supplementary Appendix 3: Description of data resources used in the linkage

Supplementary Table 1: Description of data resources used in the linkage

| Data | Population | Years collected | Data captured | Identifiers |
|------|------------|----------------|---------------|-------------|
| NPD  | All pupils in England in state schools | School Census/PLASC: 2002 onwards Early Year Census (EYC): 2008-2013 only for 3- and 4-year-olds, and 2014 onwards for all pupils Absences: 2006 onwards (exclusions, 2002 onwards) Attainment: EYFSP: 2003–2006 for only 10% of pupils, 2007 onwards for all pupils KS1: 1998 onwards KS2: 1996 onwards KS3: 1998-2013 KS4: 2000 onwards KS5: 2001 onwards | Information about all pupils who are currently at school from reception to progression at each key stage. Pupil level data on demographic and personal details, school attended, as well as whether the pupil receives support for special educational needs and/or eligibility for receiving free school meals. Pupil level data on absences and exclusions per term. Pupil level data on attainment at the EYFSP (ages 3 to 5) and Key Stages 1 to 5 (ages 7 to 18). | First name(s), surname(s), date of birth, sex, residence postcode history, residence postcodes dates entered each term. |
| HES  | All activity in English hospitals | APC (Inpatients): April 1997 onwards* HES-ONS link mortality: 1997 onwards Outpatients: 2003 onwards A&E: 2010 onwards** | Episode level data on all inpatient and day case discharges; information relating to admissions such as admission type, date, reason; clinical information such as diagnosis and procedure codes and demographic data. Deaths records of patients in English hospitals. Information on type of outpatient appointment, the main speciality, treatment speciality, referral source, waiting times, diagnosis and procedures. Records of each A&E attendance, time and method of arrival and departure, time spent in A&E. | First name(s), surname(s), date of birth, sex, residence postcode, NHS number entered for each NHS contact. |
| PDS  | All NHS patients | 2004 onwards | Demographic data of users of health and care services in England. | First name(s), surname(s), date of birth, sex, residence postcodes history, residence postcodes dates, address, NHS number entered each time GP registration is updated, or a hospital enters a different address. |

Notes: * APC (Inpatient) data has been collected since 1989 onwards. However, it can be linked only from 1997 due to the introduction of NHS numbers, which is an important element in the linkage of data. ** A&E collection was first started in 2007 but the coverage and quality improved from 2010 onwards.

NPD National Pupil Dataset; PDS Personal Demographics Service; HES Hospital Episode Statistics; NHS National Health Service; ONS Office for National Statistics; APC Admitted Patient Care; A&E Accident and Emergency; EYC Early Year Census; EYFSP Early years foundation stage profile; KS Key Stage.
**Supplementary Table 2: Example of NPD linkage table**

| aPMR | Cohort | First name | Surname | Date of birth | Sex | Postcode | Year* |
|------|--------|------------|---------|---------------|-----|----------|-------|
| 1    | 1      | John       | Smith   | 01/01/1988    | 1   | LS0 0AA  | 2004  |
| 1    | 1      | John       | Smith   | 01/01/1988    | 1   | LS1 1AA  | 2005  |
| 1    | 1      | John       | Jones   | 01/01/1988    | 1   | LS2 2AA  | 2006  |

Notes: All records for all pupils in all four cohorts were combined into a single table for linking (multiple rows for each pupil) * Year refers to academic year (e.g. 2004 means 01/09/2003 – 31/08/2004, and so on for all subsequent years).

**Supplementary Table 3: Example of PDS linkage table**

| NHS Number | First name | Surname | Date of birth | Sex | Postcode | P/A/N Start date | P/A/N End date |
|------------|------------|---------|---------------|-----|----------|------------------|----------------|
| 123456     | John       | Smith   | 01/01/1988    | 1   | LS0 0AA  | 01/01/2004       | NULL           |
| 123456     | John       | Smith   | 01/01/1988    | 1   | LS1 1AA  | 01/01/2005       | NULL           |
| 123456     | John       | Jones   | 01/01/1988    | 1   | LS2 2AA  | 01/01/2006       | NULL           |

Notes: All records from PDS table for person, name and address were combined into a single table for linking (multiple rows for each NHS Number). NULL end dates mean that the record is current.
Supplementary Appendix 4: Description of linkage between Personal Demographics Service and National Pupil Dataset

Using NPD census extracts, NHS Digital created the linkage table with identifiers for pupils in all four cohorts (Supplementary Table 2). The linkage table contained multiple rows for each pupil indicating different school censuses. For a link to be found in the linkage stage 1, besides considering the 8 steps in Table 1, the academic year from the NPD (column Year in Supplementary Table 2) needed to overlap the date intervals recorded in PDS records corresponding the period when specific demographic information was updated (see paragraph below).

Using PDS, NHS Digital created a table with patients’ identifiers (Supplementary Table 3). In this table, multiple rows for each patient’s NHS number were available referring to dates when specific patient demographic information was updated (date of birth, sex, address and name). For each NHS Number, when a record is updated, the date is recorded in the P/A/N Start date and the P/A/N End date variables. These two dates define intervals that we used in the linkage algorithm. An interval corresponds to the period between the P/A/N Start date and the P/A/N End date. P refers to a PDS record resulting from a person change (date of birth and sex), A refers to address change, and N refers to name change.

The linking algorithm takes into account all three PDS date types: Person (P), Name (N) and Address (A). The end date is the date P/A/N changes ceases to be applicable. The start date is the date that P/A/N changes start to be applicable. The two agreement dates in PDS, which are the starting and ending date of the academic year, ensure that the recent record is being linked from NPD to PDS.

Criteria for agreement dates for NPD to PDS linkage

The academic year from the NPD needs to overlap the PDS date intervals for all the relevant types, i.e. Person, Name, and Address. For example, for a link to be found when looking at an NPD record for spring census 2004, the relevant data fields (i.e. a combination of name, date of birth, sex and postcode) must link and the following criterion must hold:

\[(01/09/2003 <= P \text{ end date}) \land (P \text{ start date} <= 31/08/2004) \land (01/09/2003 <= A \text{ end date}) \land (A \text{ start date} <= 31/08/2004) \land (01/09/2003 <= N \text{ end date}) \land (N \text{ start date} <= 31/08/2004)\]

This criterion means that the academic intervals created by P/A/N Start date and End date must be within the corresponding academic year. In Panels A and B of Supplementary Figure 2 we show examples in which the previous criteria hold.
Supplementary Figure S2: Criteria for agreement dates, example using NPD spring census 2004 and three types of PDS dates

Panel A: Agreement – Example 1

Panel B: Agreement – Example 2

Note: This figure displays P/A/N dates for the academic years 2003/04. In this example, the agreement dates in PDS are 01/09/2003 and 31/08/2004. P = Person, N = Name, A = Address.
Supplementary Appendix 5: Description of demographic variables in national pupil dataset

Region 2011 pupil’s residence
We derived the region of pupil’s residence using the National Statistics Postcode Directory Lower Layer Super Output Area (LSOA11) derived from the pupil’s postcode (based on 2011 Census). We used the first available LSOA11 in the spring censuses.

Ethnic group
We derived this variable as the most common recorded ethnic group in the National Pupil Database. We used ethnicgroup variable before 2006 and ethnicgroupmajor_spr from 2006 onwards. We derived six categories to classify ethnic groups (White, Asian, Black, Chinese, Any other ethnic group, and Mixed). The following tables describe the codes assigned to each ethnic category.

Sex and IDACI deciles
We used the gender variable recorded and Income Deprivation Affecting Children Indices (IDACI) deciles derived from the pupil’s postcode (IDACIRank). We used the first available IDACI decile in the Spring censuses.

Age at start academic year
We used the age at the start academic year (AgeAtStartOfAcademicYear) that was recorded in the first spring census available.

Persistent absence
Defined as whether a child was absent in 10% or more of academic sessions.
We defined the authorised annual absence rate for all academic years available as the number of sessions missed due to authorised absence during the academic year (AuthorisedAbsence_3Term_ab) divided by the number of sessions possible for the academic year (SessionsPossible_3Term_ab). Numerator and denominator are based on all schools’ termly sessions except special schools for which annual sessions’ data are used.

Special Education Need (SEN); School Action or Early Years Action, School Action Plus or Early Years Action and SEN support (AAP/S); and Statement and Education, health and care plan (S/EHCP)
We derived the SEN, AAP/S and S/EHCP variables using the classification in the following table.
Supplementary Table 4.1: Classification of ethnic group with the *ethnicgroup* variable

| ethnicgroup                  | Ethnic group | Code |
|-----------------------------|--------------|------|
| Any Other White Background  | White        | 1    |
| Gypsy / Romany              | White        | 1    |
| Irish                       | White        | 1    |
| Traveller Of Irish Heritage | White        | 1    |
| White British               | White        | 1    |
| Any Other Asian Background  | Asian        | 2    |
| Bangladeshi                 | Asian        | 2    |
| Indian                      | Asian        | 2    |
| Pakistani                   | Asian        | 2    |
| African                     | Black        | 3    |
| Any Other Black Background  | Black        | 3    |
| Caribbean                   | Black        | 3    |
| Chinese                     | Chinese      | 4    |
| Any Other Ethnic Group      | Any other ethnic group | 5 |
| Any Other Mixed Background  | Mixed        | 6    |
| White and Asian             | Mixed        | 6    |
| White and Black African     | Mixed        | 6    |
| White and Black Caribbean   | Mixed        | 6    |
| Information Not Obtained    | Missing      | 7    |
| Missing                     | Missing      | 7    |
| Refused                     | Missing      | 7    |

Supplementary Table 4.2: Classification of ethnic group with the *ethnicgroupmajor* variable

| ethnicgroupmajor            | Ethnic group | Code |
|-----------------------------|--------------|------|
| WHIT                        | White        | 1    |
| ASIA                        | Asian        | 2    |
| BLAC                        | Black        | 3    |
| CHIN                        | Chinese      | 4    |
| AOEG                        | Any Other Ethnic Group | 5 |
| MIXD                        | Mixed        | 6    |
| UNCL                        | Missing      | 7    |

Supplementary Table 4.3: Classification of SEN, AAP/S and S/EHCP variables

| Description                          | Code | SEN: special education need | AAP/S: School action or early years action, school action plus or early years action and SEN support | S/EHCP: statement and education, health and care plan |
|--------------------------------------|------|-----------------------------|-------------------------------------------------------------------------------------------------|-----------------------------------------------------|
| No Special Educational Need          | N    |                            |                                                                                               |                                                     |
| School Action or Early Years Action (up to 2014/15) | A    | x                           |                                                                                               |                                                     |
| School Action Plus or Early Years Action Plus (up to 2014/15) | P    | x                           | x                                                                                              |                                                     |
| SEN support (since 2014/15)          | K    | x                           | x                                                                                              |                                                     |
| Statement (up to 2017/18)            | S    | x                           | x                                                                                              |                                                     |
| Education, health and care plan (since 2014/15) | E    | x                           | x                                                                                              |                                                    |
Supplementary Appendix 6. Performance of linkage stages

### Supplementary Table 5.1: Pupils records by linkage stage (stages 1 and 2) and cohort

| Cohort | Linked NPD-PDS-HES (unique aPMR) | Excluded as a fail to link PDS (stage 1) | Excluded as a fail to link HES (stage 2) | Total (unique aPMR) |
|--------|----------------------------------|----------------------------------------|----------------------------------------|---------------------|
| 1      | 571,713                          | 20,244                                 | 20,815                                 | 612,772             |
| 2      | 542,035                          | 8,868                                  | 19,839                                 | 570,742             |
| 3      | 512,526                          | 683                                    | 14,858                                 | 528,067             |
| 4      | 569,851                          | 528                                    | 5,711                                  | 576,090             |
| Total  | 2,196,125                        | 30,323                                 | 61,223                                 | 2,287,671           |

**Row percentages**

| Cohort | NPD-PDS-HES | Excluded as a fail to link PDS (stage 1) | Excluded as a fail to link HES (stage 2) | Total (all pairs) |
|--------|-------------|----------------------------------------|----------------------------------------|---------------------|
| 1      | 93.3        | 3.3                                    | 3.4                                    | 100.0               |
| 2      | 95.0        | 1.6                                    | 3.5                                    | 100.0               |
| 3      | 97.1        | 0.1                                    | 2.8                                    | 100.0               |
| 4      | 98.9        | 0.1                                    | 1.0                                    | 100.0               |
| Total  | 96.0        | 1.3                                    | 2.7                                    | 100.0               |

Notes: NPD = national pupil dataset; PDS = personal demographics service; HES = hospital episode statistics.

### Supplementary Table 5.2: Pupils records by linkage stage (stages 1 and 2), merging stage and cohort

| Cohort | Linked NPD-PDS-HES (all pairs) | Excluded as a fail to link PDS (stage 1) | Excluded as a fail to link HES (stage 2) | Not merged to UCL-HES (merging stage) | Total (all pairs) |
|--------|---------------------------------|----------------------------------------|----------------------------------------|--------------------------------------|---------------------|
| 1      | 572,673                         | 20,244                                 | 20,815                                 | 6,782                                | 613,732             |
| 2      | 543,211                         | 8,868                                  | 19,839                                 | 6,425                                | 571,918             |
| 3      | 514,369                         | 683                                    | 14,858                                 | 6,311                                | 529,910             |
| 4      | 572,570                         | 528                                    | 5,711                                  | 2,006                                | 578,809             |
| Total  | 2,202,823                       | 30,323                                 | 61,223                                 | 21,524                               | 2,294,369           |

**Row percentages**

| Cohort | NPD-HES | Excluded as a fail to link PDS (stage 1) | Excluded as a fail to link HES (stage 2) | Not merged to UCL-HES (merging stage) | Total (unique aPMR) |
|--------|---------|----------------------------------------|----------------------------------------|--------------------------------------|---------------------|
| 1      | 92.2    | 3.3                                    | 3.4                                    | 1.1                                  | 100.0               |
| 2      | 93.8    | 1.6                                    | 3.5                                    | 1.1                                  | 100.0               |
| 3      | 95.9    | 0.1                                    | 2.8                                    | 1.2                                  | 100.0               |
| 4      | 98.6    | 0.1                                    | 1.0                                    | 0.3                                  | 100.0               |
| Total  | 95.1    | 1.3                                    | 2.7                                    | 0.9                                  | 100.0               |

Notes: NPD = national pupil dataset; PDS = personal demographics service; HES = hospital episode statistics.
Supplementary Table 6.1: Linkage national pupil dataset to personal demographics service by linkage step and cohort (all pairs)

| Linkage step | Cohort (N) | Cohort (%) |
|--------------|------------|------------|
|              | 90/91      | 96/97      | 99/00      | 04/05 | Total | 90/91 | 96/97 | 99/00 | 04/05 | Total |
| 1            | 541,397    | 539,166    | 509,884    | 548,099 | 2,138,546 | 91.2 | 95.8 | 96.4 | 94.8 | 94.5 |
| 2            | 14,968     | 10,304     | 9,007      | 12,564  | 46,843    | 2.5  | 1.7  | 1.7  | 2.2  | 2.1  |
| 3            | 3,945      | 2,226      | 1,849      | 2,257   | 10,277    | 0.7  | 0.4  | 0.4  | 0.4  | 0.5  |
| 4            | 173        | 107        | 177        | 503     | 960       | 0.0  | 0.0  | 0.0  | 0.1  | 0.0  |
| 5            | 8,532      | 4,835      | 3,887      | 5,617   | 22,871    | 1.4  | 0.9  | 0.7  | 1.0  | 1.0  |
| 6            | 3,582      | 1,288      | 1,016      | 2,147   | 8,033     | 0.6  | 0.2  | 0.2  | 0.4  | 0.4  |
| 7            | 19,222     | 4,694      | 3,096      | 6,510   | 33,522    | 3.2  | 0.8  | 0.6  | 1.1  | 1.5  |
| 8            | 1,669      | 430        | 311        | 584     | 2,994     | 0.3  | 0.1  | 0.1  | 0.1  | 0.1  |
| Total        | 593,488    | 563,050    | 529,227    | 578,281 | 2,264,046 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |

Note: The table includes successful linked records.

Supplementary Table 6.2: Linkage national pupil dataset to personal demographics service by academic year of linkage step and cohort (all pairs)

| Academic year of linkage step | Cohort (N) | Cohort (%) |
|-------------------------------|------------|------------|
|                               | 90/91      | 96/97      | 99/00      | 04/05 | Total | 90/91 | 96/97 | 99/00 | 04/05 | Total |
| 2004                          | 503,135    | 441,016    | 179,013    | 0     | 1,123,164 | 84.8 | 78.3 | 33.8 | 0.0  | 49.6 |
| 2005                          | 34,458     | 40,140     | 262,752    | 0     | 337,351    | 5.8  | 7.1  | 49.7 | 0.0  | 14.9 |
| 2006                          | 16,444     | 18,744     | 23,689     | 0     | 58,877     | 2.8  | 3.3  | 4.5  | 0.0  | 2.6  |
| 2007                          | 12,676     | 14,434     | 15,735     | 0     | 42,846     | 2.1  | 2.6  | 3.0  | 0.0  | 1.9  |
| 2008                          | 4,312      | 11,561     | 11,246     | 0     | 27,119     | 0.7  | 2.1  | 2.1  | 0.0  | 1.2  |
| 2009                          | 1,572      | 18,223     | 7,369      | 228,484 | 255,646    | 0.3  | 3.2  | 1.4  | 39.5 | 11.3 |
| 2010                          | 0          | 5,047      | 5,825      | 292,282 | 303,154    | 0.0  | 0.9  | 1.1  | 50.5 | 13.4 |
| 2011                          | 0          | 3,423      | 5,579      | 18,952  | 27,954     | 0.0  | 0.6  | 1.1  | 3.3  | 1.2  |
| 2012                          | 0          | 2,046      | 6,137      | 6,688   | 14,871     | 0.0  | 0.4  | 1.2  | 1.2  | 0.7  |
| 2013                          | 0          | 1,725      | 2,394      | 9,397   | 13,516     | 0.0  | 0.3  | 0.5  | 1.6  | 0.6  |
| 2014                          | 0          | 1,032      | 1,828      | 4,542   | 7,402      | 0.0  | 0.2  | 0.4  | 0.8  | 0.3  |
| 2015                          | 0          | 535        | 2,382      | 5,324   | 8,441      | 0.0  | 0.1  | 0.5  | 0.9  | 0.4  |
| 2016                          | 0          | 0          | 1,475      | 4,361   | 5,836      | 0.0  | 0.0  | 0.3  | 0.8  | 0.3  |
| 2017                          | 0          | 0          | 128        | 879     | 1,007      | 0.0  | 0.0  | 0.0  | 0.2  | 0.0  |
| 2018                          | 0          | 0          | 68         | 278     | 346        | 0.0  | 0.0  | 0.0  | 0.1  | 0.0  |
| 9999                          | 20,891     | 5,124      | 3,407      | 7,094   | 36,516     | 3.5  | 0.9  | 0.6  | 1.2  | 1.6  |
| Total                         | 593,488    | 563,050    | 529,227    | 578,281 | 2,264,046  | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |

Note: The table includes successful linked records.
Supplementary Table 6.3: Linkage national pupil dataset to personal demographics service by linkage step and ethnicity (all pairs)

| Linkage step | White  | Asian  | Black  | Chinese | Any other ethnic group | Mixed | Missing | Total  |
|--------------|--------|--------|--------|---------|------------------------|-------|---------|--------|
| Records (N)  | 1,783,487 | 157,904 | 80,501 | 6,196   | 19,298                | 80,098 | 11,062  | 2,138,546 |
| 2            | 28,274  | 8,204  | 5,393  | 216     | 1,609                 | 2,781  | 366     | 46,843 |
| 3            | 6,400   | 2,006  | 1,009  | 113     | 274                   | 383    | 92      | 10,277 |
| 4            | 432     | 227    | 172    | 0       | 53                    | 57     | 19      | 960    |
| 5            | 9,281   | 7,273  | 3,664  | 313     | 862                   | 1,285  | 193     | 22,871 |
| 6            | 5,557   | 988    | 753    | 36      | 230                   | 376    | 93      | 8,033  |
| 7            | 26,620  | 2,355  | 2,022  | 169     | 418                   | 1,361  | 577     | 33,522 |
| 8            | 1,704   | 430    | 459    | 14      | 134                   | 176    | 77      | 2,994  |
| Total        | 1,861,755 | 179,387 | 93,973 | 7,057   | 22,878                | 86,517 | 12,479  | 2,264,046 |

Row percentages

| Linkage step | White | Asian | Black | Chinese | Any other ethnic group | Mixed | Missing | Total |
|--------------|-------|-------|-------|---------|------------------------|-------|---------|-------|
| 1            | 83.4  | 7.4   | 3.8   | 0.3     | 0.9                    | 3.7   | 0.5     | 100.0 |
| 2            | 60.4  | 17.5  | 11.5  | 0.5     | 3.4                    | 5.9   | 0.8     | 100.0 |
| 3            | 62.3  | 19.5  | 9.8   | 1.1     | 2.7                    | 3.7   | 0.9     | 100.0 |
| 4            | 45.0  | 23.6  | 17.9  | 0.0     | 5.5                    | 5.9   | 2.0     | 100.0 |
| 5            | 40.6  | 31.8  | 16.0  | 1.4     | 3.8                    | 5.6   | 0.8     | 100.0 |
| 6            | 69.2  | 12.3  | 9.4   | 0.4     | 2.9                    | 4.7   | 1.2     | 100.0 |
| 7            | 79.4  | 7.0   | 6.0   | 0.5     | 1.2                    | 4.1   | 1.7     | 100.0 |
| 8            | 56.9  | 14.4  | 15.3  | 0.5     | 4.5                    | 5.9   | 2.6     | 100.0 |
| Total        | 82.2  | 7.9   | 4.2   | 0.3     | 1.0                    | 3.8   | 0.6     | 100.0 |

Column percentages

| Linkage step | White | Asian | Black | Chinese | Any other ethnic group | Mixed | Missing | Total |
|--------------|-------|-------|-------|---------|------------------------|-------|---------|-------|
| 1            | 95.8  | 88.0  | 85.7  | 87.8    | 84.4                   | 92.6  | 88.6    | 94.5  |
| 2            | 1.5   | 4.6   | 5.7   | 3.1     | 7.0                    | 3.2   | 2.9     | 2.1   |
| 3            | 0.3   | 1.1   | 1.1   | 1.6     | 1.2                    | 0.4   | 0.7     | 0.5   |
| 4            | 0.0   | 0.1   | 0.2   | 0.0     | 0.2                    | 0.1   | 0.2     | 0.0   |
| 5            | 0.5   | 4.1   | 3.9   | 4.4     | 3.8                    | 1.5   | 1.5     | 1.0   |
| 6            | 0.3   | 0.6   | 0.8   | 0.5     | 1.0                    | 0.4   | 0.7     | 0.4   |
| 7            | 1.4   | 1.3   | 2.2   | 2.4     | 1.8                    | 1.6   | 4.6     | 1.5   |
| 8            | 0.1   | 0.2   | 0.5   | 0.2     | 0.6                    | 0.2   | 0.6     | 0.1   |
| Total        | 100.0 | 100.0 | 100.0 | 100.0   | 100.0                  | 100.0 | 100.0   | 100.0 |

Note: The table includes successful linked records.
Supplementary Table 6.4: Linkage national pupil dataset to personal demographics service by linkage step and region (all pairs)

| Linkage step | London | South East | South West | West Midlands | North West | North East | Yorkshire and The Humber | East Midlands | East of England | Wales | Missing | Total |
|--------------|--------|------------|------------|---------------|------------|-----------|--------------------------|--------------|-----------------|-------|----------|-------|
| Records (n)  | 278,298| 315,432    | 196,505    | 233,888       | 297,522    | 104,861   | 213,681                  | 175,180      | 225,551         | 97,435| 2,138,546|       |
| 1            | 11,609 | 5,149      | 2,934      | 5,004         | 6,838      | 1,861     | 4,449                    | 2,893        | 3,406           | 0     | 2,700    | 46,843|
| 2            | 2,673  | 1,173      | 740        | 965           | 1,435      | 386       | 863                      | 609          | 858             | 0     | 575      | 10,277|
| 3            | 340    | 64         | 35         | 130           | 121        | 28        | 90                       | 33           | 62              | 0     | 57       | 960   |
| 4            | 6,599  | 1,976      | 901        | 2,489         | 3,503      | 831       | 2,374                    | 1,205        | 1,421           | 0     | 1,572    | 22,871|
| 5            | 1,766  | 876        | 531        | 930           | 1,336      | 352       | 701                      | 439          | 631             | 0     | 471      | 8,033 |
| 6            | 5,770  | 4,424      | 2,726      | 3,658         | 5,611      | 1,367     | 2,728                    | 2,045        | 2,970           | 21    | 2,202    | 33,522|
| 7            | 871    | 267        | 151        | 266           | 464        | 78        | 214                      | 135          | 273             | 0     | 275      | 2,994 |
| Total        | 307,926| 329,361    | 204,523    | 247,330       | 316,830    | 109,764   | 225,100                  | 182,539      | 235,172         | 105,287| 2,264,046|       |

Supplementary Table 7.1: Linkage personal demographics service to hospital episode statistics by linkage step and cohort (all pairs)

| Cohort (N) | Linkage step | 90/91 | 96/97 | 99/00 | 04/05 | Total |
|------------|---------------|-------|-------|-------|-------|-------|
| 1          | 455,015       | 435,678| 412,723| 488,064| 1,791,480|
| 2          | 112,446       | 100,895| 92,854| 80,384| 386,579|
| 3          | 2,080         | 863    | 727    | 1,705  | 5,375  |
| 4          | 656           | 305    | 269    | 443    | 1,673  |
| 5          | 179           | 190    | 234    | 515    | 1,118  |
| 6          | 2,181         | 5,059  | 7,314  | 1,266  | 15,820 |
| 7          | 47            | 97     | 135    | 24     | 303    |
| 8          | 69            | 124    | 113    | 169    | 475    |
| Total      | 572,673       | 543,211| 514,369| 572,570| 2,202,823|

Note: The table includes successful linked records.
Supplementary Table 7.2: Linkage personal demographics service to hospital episode statistics by linkage step and ethnicity (all pairs)

| Linkage step | White (n) | Asian (n) | Black (n) | Chinese (n) | Any other ethnic group (n) | Mixed (n) | Missing (n) | Total (n) |
|--------------|-----------|-----------|-----------|-------------|---------------------------|-----------|-------------|-----------|
| 1            | 1,481,539 | 144,364   | 68,528    | 5,055       | 16,422                    | 66,366    | 9,206       | 1,791,480 |
| 2            | 315,316   | 26,125    | 20,160    | 1,323       | 4,508                     | 16,671    | 2,476       | 386,579   |
| 3            | 3,716     | 689       | 508       | 32          | 128                       | 246       | 56          | 5,375     |
| 4            | 1,119     | 205       | 194       | 0           | 52                        | 89        | 14          | 1,673     |
| 5            | 660       | 195       | 138       | 0           | 47                        | 66        | 12          | 1,118     |
| 6            | 11,333    | 2,083     | 1,268     | 114         | 302                       | 626       | 94          | 15,820    |
| 7            | 233       | 33        | 18        | 0           | 0                         | 19        | 0           | 303       |
| 8            | 300       | 69        | 59        | 0           | 15                        | 32        | 0           | 475       |
| Total        | 1,814,216 | 173,763   | 90,873    | 6,524       | 21,474                    | 84,115    | 11,858      | 2,202,823 |

Row percentages

| Linkage step | White | Asian | Black | Chinese | Any other ethnic group | Mixed | Missing | Total |
|--------------|-------|-------|-------|---------|------------------------|-------|---------|-------|
| 1            | 82.7  | 8.1   | 3.8   | 0.3     | 0.9                    | 3.7   | 0.5     | 100.0 |
| 2            | 81.6  | 6.8   | 5.2   | 0.6     | 0.9                    | 4.3   | 0.6     | 100.0 |
| 3            | 69.1  | 12.8  | 9.5   | 0.6     | 3.1                    | 4.6   | 1.0     | 100.0 |
| 4            | 66.9  | 12.3  | 11.6  | 0.0     | 3.1                    | 5.3   | 0.8     | 100.0 |
| 5            | 59.0  | 17.4  | 12.3  | 0.0     | 4.2                    | 5.9   | 1.1     | 100.0 |
| 6            | 71.6  | 13.2  | 8.0   | 0.7     | 1.9                    | 4.0   | 0.6     | 100.0 |
| 7            | 76.9  | 10.9  | 5.9   | 0.0     | 0.0                    | 6.3   | 0.0     | 100.0 |
| 8            | 63.2  | 14.5  | 12.4  | 0.0     | 3.2                    | 6.7   | 0.0     | 100.0 |
| Total        | 82.4  | 7.9   | 4.1   | 0.3     | 1.0                    | 3.8   | 0.5     | 100.0 |

Column percentages

| Linkage step | White | Asian | Black | Chinese | Any other ethnic group | Mixed | Missing | Total |
|--------------|-------|-------|-------|---------|------------------------|-------|---------|-------|
| 1            | 81.7  | 83.1  | 75.4  | 77.5     | 76.5                    | 78.9  | 77.6    | 81.3  |
| 2            | 17.4  | 15.0  | 22.2  | 20.3     | 21.0                    | 19.8  | 20.9    | 17.5  |
| 3            | 0.2   | 0.4   | 0.6   | 0.5      | 0.6                    | 0.3   | 0.5     | 0.2   |
| 4            | 0.1   | 0.1   | 0.2   | 0.0      | 0.2                    | 0.1   | 0.1     | 0.1   |
| 5            | 0.0   | 0.1   | 0.2   | 0.0      | 0.2                    | 0.1   | 0.1     | 0.1   |
| 6            | 0.6   | 1.2   | 1.4   | 1.7      | 1.4                    | 0.7   | 0.8     | 0.7   |
| 7            | 0.0   | 0.0   | 0.0   | 0.0      | 0.0                    | 0.0   | 0.0     | 0.0   |
| 8            | 0.0   | 0.0   | 0.1   | 0.0      | 0.1                    | 0.0   | 0.0     | 0.0   |
| Total        | 100.0 | 100.0 | 100.0 | 100.0    | 100.0                  | 100.0 | 100.0   | 100.0 |

Note: The table includes successful linked records.
**Supplementary Table 7.3: Linkage personal demographics service to hospital episode statistics by linkage step and region (all pairs)**

| Linkage step | London | South East | South West | West Midlands | North West | North East | Yorkshire and The Humber | East Midlands | East of England | Wales | Missing | Total |
|--------------|--------|------------|------------|---------------|------------|------------|--------------------------|---------------|----------------|-------|----------|-------|
| Records (n)  |        |            |            |               |            |            |                          |               |                |       |          |       |
| 1            | 278,298| 315,432    | 196,505    | 233,888       | 297,522    | 104,861    | 213,681                  | 175,180       | 225,551        | 193   | 97,435   | 2,138,546 |
| 2            | 11,609 | 5,149      | 2,934      | 6,838         | 1,861      | 4,449      | 2,893                    | 3,406         | 0              | 2,700 | 46,843   |         |
| 3            | 2,673  | 1,173      | 740        | 965           | 1,435      | 386        | 863                      | 609           | 858            | 0     | 57       | 10,277  |
| 4            | 340    | 64         | 35         | 130           | 121        | 28         | 90                       | 33            | 62             | 0     | 57       | 960     |
| 5            | 6,599  | 1,976      | 901        | 2,489         | 1,336      | 701        | 1,205                    | 1,421         | 0              | 1,572 | 22,871   |         |
| 6            | 1,766  | 876        | 531        | 930           | 352        | 701        | 439                      | 631           | 0              | 471   | 8,033    |         |
| 7            | 5,770  | 4,424      | 2,726      | 3,658         | 5,611      | 2,728      | 2,045                    | 2,970         | 21             | 2,202 | 33,522   |         |
| 8            | 871    | 267        | 151        | 266           | 464        | 78         | 214                      | 135           | 273            | 0     | 275      | 2,994   |
| Total        | 307,926| 329,361    | 204,523    | 247,330       | 316,830    | 109,764    | 225,100                  | 182,539       | 235,172        | 221   | 105,280  | 2,264,046|

**Row percentages**

| Linkage step | London | South East | South West | West Midlands | North West | North East | Yorkshire and The Humber | East Midlands | East of England | Wales | Missing | Total |
|--------------|--------|------------|------------|---------------|------------|------------|--------------------------|---------------|----------------|-------|----------|-------|
| 1            | 13.0   | 14.7       | 9.2        | 10.9          | 4.9        | 10.0       | 8.2                      | 10.5          | 0.0            | 4.6   | 100.0    |       |
| 2            | 24.8   | 11.0       | 6.3        | 10.7          | 4.0        | 9.5        | 6.2                      | 7.3           | 0.0            | 5.8   | 100.0    |       |
| 3            | 26.0   | 11.4       | 7.2        | 9.4           | 14.0       | 8.4        | 9.4                      | 5.9           | 8.3            | 0.6   | 100.0    |       |
| 4            | 35.4   | 6.7        | 3.6        | 13.5          | 12.6       | 9.4        | 4.8                      | 10.5          | 2.9            | 6.9   | 100.0    |       |
| 5            | 28.9   | 8.6        | 3.9        | 10.9          | 15.3       | 10.4       | 9.6                      | 7.9           | 6.2            | 0.9   | 100.0    |       |
| 6            | 22.0   | 10.9       | 6.6        | 11.6          | 16.6       | 8.7        | 7.5                      | 3.9           | 8.9            | 0.4   | 100.0    |       |
| 7            | 17.2   | 13.2       | 8.1        | 10.9          | 16.7       | 8.1        | 7.1                      | 4.1           | 9.2            | 0.1   | 100.0    |       |
| 8            | 29.1   | 8.9        | 5.0        | 8.9           | 15.5       | 7.1        | 5.1                      | 10.4          | 0.0            | 4.7   | 100.0    |       |
| Total        | 13.6   | 14.5       | 9.0        | 10.9          | 14.0       | 9.9        | 8.1                      | 10.4          | 0.0            |       | 100.0    |       |

**Column percentages**

| Linkage step | London | South East | South West | West Midlands | North West | North East | Yorkshire and The Humber | East Midlands | East of England | Wales | Missing | Total |
|--------------|--------|------------|------------|---------------|------------|------------|--------------------------|---------------|----------------|-------|----------|-------|
| 1            | 90.4   | 95.8       | 96.1       | 94.6          | 93.9       | 95.5       | 94.9                     | 96.0          | 95.9           | 87.3  | 92.5     | 94.5  |
| 2            | 3.8    | 1.6        | 1.4        | 2.0           | 2.2        | 1.7        | 2.0                      | 1.6           | 1.4            | 0.0   | 2.6      | 2.1   |
| 3            | 0.9    | 0.4        | 0.4        | 0.4           | 0.5        | 0.4        | 0.3                      | 0.3           | 0.4            | 0.0   | 0.5      | 0.5   |
| 4            | 0.1    | 0.0        | 0.0        | 0.1           | 0.0        | 0.0        | 0.0                      | 0.0           | 0.0            | 0.1   | 0.0      | 0.1   |
| 5            | 2.1    | 0.6        | 0.4        | 1.0           | 1.1        | 0.8        | 1.1                      | 0.7           | 0.6            | 0.0   | 1.5      | 1.0   |
| 6            | 0.6    | 0.3        | 0.3        | 0.4           | 0.4        | 0.3        | 0.3                      | 0.2           | 0.3            | 0.0   | 0.4      | 0.4   |
| 7            | 1.9    | 1.3        | 1.3        | 1.5           | 1.8        | 1.2        | 1.2                      | 1.1           | 1.3            | 9.5   | 2.1      | 1.5   |
| 8            | 0.3    | 0.1        | 0.1        | 0.1           | 0.1        | 0.1        | 0.1                      | 0.1           | 0.1            | 0.0   | 0.3      | 0.1   |
| Total        | 100.0  | 100.0      | 100.0      | 100.0         | 100.0      | 100.0      | 100.0                    | 100.0         | 100.0          | 100.0 | 100.0    | 100.0 |

Note: The table includes successful linked records.
Supplementary Table 8.1: Linking rate by region 2011 pupil’s residence (first recorded) and cohort

| Region | Cohort 1990/91 | Cohort 1996/97 | Cohort 1999/00 | Cohort 2004/05 |
|--------|----------------|----------------|----------------|----------------|
| London | Unlinked to HES | Linked to HES | Total | Unlinked to HES | Linked to HES | Total | Unlinked to HES | Linked to HES | Total | Unlinked to HES | Linked to HES | Total |
|        | 7,708 | 67,898 | 75,606 | 6,205 | 71,463 | 77,668 | 4,214 | 70,706 | 74,920 | 1,552 | 83,385 | 84,937 |
| South East | 7,985 | 81,697 | 89,682 | 5,942 | 75,315 | 81,257 | 3,848 | 73,988 | 77,836 | 1,324 | 83,462 | 84,786 |
| South West | 4,212 | 51,970 | 56,182 | 3,014 | 50,234 | 53,248 | 2,008 | 70,358 | 72,366 | 953 | 75,962 | 76,915 |
| West Midlands | 4,906 | 62,935 | 67,841 | 3,371 | 59,467 | 62,838 | 2,233 | 54,985 | 57,218 | 735 | 60,099 | 60,834 |
| North West | 6,194 | 83,206 | 89,400 | 3,623 | 77,676 | 81,299 | 2,008 | 70,358 | 72,366 | 953 | 75,962 | 76,915 |
| North East | 1,563 | 29,287 | 30,850 | 1,019 | 27,323 | 28,342 | 571 | 24,419 | 24,990 | 187 | 25,968 | 26,056 |
| Yorkshire and The Humber | 3,872 | 57,455 | 61,327 | 2,886 | 54,467 | 57,353 | 1,481 | 49,579 | 51,060 | 652 | 56,059 | 56,711 |
| East Midlands | 3,533 | 47,048 | 50,581 | 2,757 | 42,119 | 44,876 | 1,756 | 40,828 | 42,584 | 669 | 45,069 | 45,738 |
| East of England | 5,526 | 59,620 | 65,146 | 4,500 | 54,325 | 58,825 | 3,085 | 52,101 | 55,186 | 1,010 | 57,331 | 58,341 |
| Wales | 28 | 38 | 66 | 0 | 45 | 45 | 0 | 64 | 64 | 0 | 69 | 69 |
| Missing | 2,314 | 23,777 | 26,091 | 1,808 | 22,741 | 24,549 | 1,304 | 23,645 | 24,949 | 673 | 30,692 | 31,365 |
| Total | 47,841 | 564,931 | 612,772 | 35,132 | 535,610 | 570,742 | 21,852 | 506,215 | 528,067 | 8,245 | 567,845 | 576,090 |

Supplementary Table 8.2: Linking rate by ethnic group and cohort

| Ethnic Group | Cohort 1990/91 | Cohort 1996/97 | Cohort 1999/00 | Cohort 2004/05 |
|--------------|----------------|----------------|----------------|----------------|
| White        | Unlinked to HES | Linked to HES | Total | Unlinked to HES | Linked to HES | Total | Unlinked to HES | Linked to HES | Total | Unlinked to HES | Linked to HES | Total |
|              | 27,620 | 487,698 | 515,318 | 24,325 | 452,995 | 477,320 | 15,464 | 414,538 | 430,002 | 5,081 | 437,566 | 442,647 |
| Asian        | 2,535 | 32,946 | 35,481 | 2,572 | 37,555 | 40,127 | 2,548 | 42,910 | 45,458 | 1,188 | 57,524 | 58,712 |
| Black        | 1,500 | 16,967 | 18,467 | 1,416 | 19,154 | 20,570 | 1,697 | 21,416 | 23,113 | 679 | 31,483 | 32,162 |
| Chinese      | 276 | 1,379 | 1,655 | 211 | 1,437 | 1,648 | 170 | 1,527 | 1,697 | 88 | 2,034 | 2,122 |
| Any other ethnic group | 496 | 3,610 | 4,106 | 621 | 3,935 | 4,556 | 689 | 5,110 | 5,799 | 482 | 8,341 | 8,823 |
| Mixed        | 832 | 13,775 | 14,607 | 1,272 | 19,244 | 20,516 | 1,157 | 20,098 | 21,255 | 559 | 29,699 | 30,258 |
| Missing      | 14,582 | 856 | 23,138 | 4,715 | 1,290 | 6,005 | 127 | 616 | 743 | 168 | 1,198 | 1,366 |
| Total        | 47,841 | 564,931 | 612,772 | 35,132 | 535,610 | 570,742 | 21,852 | 506,215 | 528,067 | 8,245 | 567,845 | 576,090 |

Notes: HES = hospital episode statistics.
Supplementary Table 8.3: Linking rate by sex and cohort

| Cohort 1990/91 | Cohort 1996/97 | Cohort 1999/00 | Cohort 2004/05 |
|---------------|---------------|---------------|---------------|
| Records (N)   |               |               |               |
| Male          | 27,272        | 20,512        | 57            | 47,841        |
| Female        | 20,512        | 27,272        | 57            | 47,841        |
| Missing       | 57            | 57            | 57            | 57            |
| Total         | 47,841        | 564,931       | 612,772       | 576,090       |

| Unlinked to HES | Linked to HES | Total |
|-----------------|---------------|-------|
| Male            | 16,935        | 274,966 | 291,901 |
| Female          | 18,182        | 260,603 | 278,785 |
| Missing         | 15            | 41      | 56     |
| Total           | 35,132        | 535,610 | 570,742 |

| Linkage rate (%) | Male | Female | Missing |
|------------------|------|--------|---------|
| 1990/91          | 8.7  | 6.8    | 9.3     |
| 1996/97          | 5.8  | 6.5    | 6.2     |
| 1999/00          | 3.6  | 4.8    | 4.1     |
| 2004/05          | 1.2  | 1.7    | 1.4     |

Notes: HES = hospital episode statistics.

Supplementary Table 8.4: Linking rate by IDACI deciles and cohort

| Cohort 1990/91 | Cohort 1996/97 | Cohort 1999/00 | Cohort 2004/05 |
|---------------|---------------|---------------|---------------|
| Records (N)   |               |               |               |
| 1 (deprived)  | 7,297         | 4,598         | 4,250         | 47,841        |
| 2              | 5,994         | 4,300         | 4,598         | 47,841        |
| 3              | 5,406         | 4,930         | 4,598         | 47,841        |
| 4              | 4,930         | 4,598         | 4,250         | 47,841        |
| 5              | 4,598         | 4,250         | 3,842         | 47,841        |
| 6              | 4,250         | 3,842         | 3,677         | 47,841        |
| 7              | 3,842         | 3,677         | 3,502         | 47,841        |
| 8              | 3,502         | 3,502         | 3,623         | 47,841        |
| 9              | 3,623         | 3,623         | 3,722         | 47,841        |
| 10 (affluent)  | 3,722         | 3,722         | 3,842         | 47,841        |
| Missing        | 722           | 722           | 722           | 47,841        |
| Total          | 47,841        | 47,841        | 47,841        | 576,090       |

| Unlinked to HES | Linked to HES | Total |
|-----------------|---------------|-------|
| 1 (deprived)    | 4,853         | 50,469 | 55,322 |
| 2               | 4,232         | 51,065 | 55,297 |
| 3               | 3,800         | 51,603 | 55,403 |
| 4               | 3,724         | 51,655 | 55,379 |
| 5               | 3,430         | 52,267 | 55,697 |
| 6               | 3,288         | 52,386 | 55,674 |
| 7               | 2,919         | 53,228 | 56,147 |
| 8               | 2,897         | 54,166 | 57,063 |
| 9               | 2,828         | 55,644 | 58,472 |
| 10 (affluent)   | 2,684         | 56,199 | 58,883 |
| Missing         | 477           | 6,928   | 7,405   |
| Total           | 35,132        | 535,610 | 570,742 |

| Linkage rate (%) | 1 (deprived) | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 (affluent) | Missing |
|------------------|-------------|---|---|---|---|---|---|---|---|---------------|---------|
| 1990/91          | 11.9        | 9.7| 8.8| 8.0| 7.5| 7.0| 6.3| 6.2| 6.0| 6.3           | 8.9     |
| 1996/97          | 11.9        | 9.7| 8.8| 8.0| 7.5| 7.0| 6.3| 6.2| 6.0| 6.3           | 8.9     |
| 1999/00          | 11.9        | 9.7| 8.8| 8.0| 7.5| 7.0| 6.3| 6.2| 6.0| 6.3           | 8.9     |
| 2004/05          | 11.9        | 9.7| 8.8| 8.0| 7.5| 7.0| 6.3| 6.2| 6.0| 6.3           | 8.9     |

Notes: HES = hospital episode statistics; IDACI = income deprivation affecting children index.
## Supplementary Table 8.5: Linking rate by sex, ethnicity and cohort

| Cohort | Records (n) | Unlinked to HES | Linked to HES | Total |
|--------|-------------|-----------------|---------------|-------|
| 1990/91 | Male-White | 16,520 | 246,476 | 262,996 |
|         | Male-Asian  | 1,438 | 16,594 | 18,032 |
|         | Male-Chinese | 148  | 686   | 834   |
|         | Male-Any other ethnic | 287 | 1,902 | 2,189 |
|         | Male-Mixed   | 488  | 6,794 | 7,282 |
| 1996/97 | Female-White | 11,100 | 241,222 | 252,322 |
|         | Female-Asian | 1,097 | 16,352 | 17,449 |
|         | Female-Chinese | 128 | 693   | 821   |
|         | Female-Any other ethnic | 209 | 1,708 | 1,917 |
|         | Female-Mixed | 344  | 6,981 | 7,325 |
| 1999/00 | Total       | 33,259 | 556,375 | 589,634 |

| Cohort | Linkage rate (%) | Male-White | Female-White | Male-Asian | Female-Asian | Male-Chinese | Female-Chinese | Male-Any other ethnic | Female-Any other ethnic | Male-Mixed | Female-Mixed | Total |
|--------|------------------|------------|--------------|------------|--------------|--------------|----------------|----------------------|------------------------|------------|--------------|-------|
| 1990/91 | Male-White       | 6.3        | 93.7         | 100.0      | 4.8          | 95.2         | 100.0         | 3.1                  | 96.9                   | 100.0      | 1.0          | 99.0  |
|         | Male-Asian       | 8.0        | 92.0         | 100.0      | 5.5          | 94.5         | 100.0         | 4.7                  | 95.4                   | 100.0      | 1.7          | 98.4  |
|         | Male-Chinese     | 9.8        | 90.2         | 100.0      | 6.6          | 93.4         | 100.0         | 5.8                  | 94.2                   | 100.0      | 1.6          | 98.4  |
|         | Male-Any other ethnic | 17.8 | 82.3 | 100.0 | 11.4 | 88.6 | 100.0 | 10.3 | 89.7 | 100.0 | 3.3 | 96.7 | 100.0 |
|         | Male-Mixed       | 13.1       | 86.9         | 100.0      | 13.2         | 86.8         | 100.0         | 10.9                 | 89.1                   | 100.0      | 4.5          | 95.5  |
|         | Female-White     | 4.4        | 95.6         | 100.0      | 5.4          | 94.6         | 100.0         | 4.1                  | 95.9                   | 100.0      | 1.3          | 98.7  |
|         | Female-Asian     | 6.3        | 93.7         | 100.0      | 7.4          | 92.6         | 100.0         | 6.6                  | 93.4                   | 100.0      | 2.4          | 97.6  |
|         | Female-Chinese   | 6.4        | 93.6         | 100.0      | 7.2          | 92.8         | 100.0         | 9.0                  | 91.1                   | 100.0      | 2.6          | 97.4  |
|         | Female-Any other ethnic | 15.6 | 84.4 | 100.0 | 14.2 | 85.8 | 100.0 | 9.7 | 90.3 | 100.0 | 5.0 | 95.0 | 100.0 |
|         | Female-Mixed     | 10.9       | 89.1         | 100.0      | 14.1         | 85.9         | 100.0         | 12.9                 | 87.1                   | 100.0      | 6.4          | 93.6  |
|         | Total            | 5.6        | 94.4         | 100.0      | 5.4          | 94.6         | 100.0         | 4.1                  | 95.9                   | 100.0      | 1.4          | 98.6  |

Notes: HES = hospital episode statistics. Cohort totals may differ from previous tables because the table excludes observations with missing Ethnicity.
Supplementary Appendix 8. Standardized differences and P-values

Supplementary Table 9.1: Distribution of demographic variables in national pupil dataset by linking status, N = 613,732 pairs (national pupil dataset to hospital episode statistics). Cohort 1990/91

| Region 2011 pupil’s residence (first recorded) NPD | No link to HES | Linked to HES | Total | P-value | Standardized difference |
|---------------------------------------------------|---------------|--------------|-------|---------|-------------------------|
|                                                   | N (%)         | N (%)        | N (%) |         |                         |
| London                                            | 7,729         | 68,073       | 75,802| 12.4    | 0.191                   |
| South East                                        | 8,000         | 81,806       | 89,806| 14.6    |                         |
| South West                                        | 4,217         | 52,018       | 56,235| 9.2     |                         |
| West Midlands                                     | 4,915         | 63,013       | 67,928| 11.1    |                         |
| North West                                        | 6,200         | 83,376       | 89,576| 14.6    |                         |
| North East                                        | 1,567         | 29,318       | 30,885| 5.0     |                         |
| Yorkshire and The Humber                          | 3,885         | 57,539       | 61,424| 10.6    |                         |
| East Midlands                                     | 3,535         | 47,096       | 50,631| 8.2     |                         |
| East of England                                   | 5,541         | 59,686       | 65,227| 10.6    |                         |
| Wales                                             | 28            | 38           | 66    | 0.0     |                         |
| Missing                                           | 2,317         | 23,835       | 26,152| 4.3     |                         |
| Total                                             | 47,934        | 565,798      | 613,732| 100.0  |                         |

| Ethnic group (NPD)                                | No link to HES | Linked to HES | Total | P-value | Standardized difference |
|---------------------------------------------------|---------------|--------------|-------|---------|-------------------------|
|                                                   | N (%)         | N (%)        | N (%) |         |                         |
| White                                             | 27,692        | 488,330      | 516,022| 84.1    | 0.160                   |
| Asian                                             | 2,541         | 33,024       | 35,565| 5.8     |                         |
| Black                                             | 1,507         | 17,047       | 18,554| 3.0     |                         |
| Chinese                                           | 278           | 1,384        | 1,662 | 0.3     |                         |
| Any other ethnic group                            | 498           | 3,627        | 4,125 | 0.7     |                         |
| Mixed                                             | 834           | 13,808       | 14,642| 2.4     |                         |
| Missing                                           | 14,584        | 8,578        | 23,162| 3.8     |                         |
| Total                                             | 47,934        | 565,798      | 613,732| 100.0  |                         |

| Sex (NPD)                                         | No link to HES | Linked to HES | Total | P-value | Standardized difference |
|---------------------------------------------------|---------------|--------------|-------|---------|-------------------------|
|                                                   | N (%)         | N (%)        | N (%) |         |                         |
| Male                                              | 27,334        | 285,716      | 313,050| 51.0    | 0.131                   |
| Female                                            | 20,543        | 279,520      | 300,063| 48.9    |                         |
| Missing                                           | 57            | 562          | 619   | 0.1     |                         |
| Total                                             | 47,934        | 565,798      | 613,732| 100.0  |                         |

| IDACI Deciles (first Census) NPD                   | No link to HES | Linked to HES | Total | P-value | Standardized difference |
|---------------------------------------------------|---------------|--------------|-------|---------|-------------------------|
|                                                   | N (%)         | N (%)        | N (%) |         |                         |
| 1 (deprived)                                      | 7,306         | 54,336       | 61,642| 10.0    | 0.242                   |
| 2                                                  | 6,001         | 55,606       | 61,607| 10.0    |                         |
| 3                                                  | 5,414         | 56,149       | 61,563| 10.0    |                         |
| 4                                                  | 4,941         | 56,600       | 61,541| 10.0    |                         |
| 5                                                  | 4,611         | 56,620       | 61,231| 10.0    |                         |
| 6                                                  | 4,255         | 56,927       | 61,182| 10.0    |                         |
| 7                                                  | 3,854         | 56,891       | 60,745| 9.9     |                         |
| 8                                                  | 3,685         | 56,122       | 59,807| 9.7     |                         |
| 9                                                  | 3,514         | 54,875       | 58,389| 9.5     |                         |
| 10 (affluent)                                     | 3,630         | 54,286       | 57,916| 9.4     |                         |
| Missing                                           | 723           | 7,386        | 8,109 | 1.3     |                         |
| Total                                             | 47,934        | 565,798      | 613,732| 100.0  |                         |

| Age at start academic year (first recorded) NPD    | No link to HES | Linked to HES | Total | P-value | Standardized difference |
|---------------------------------------------------|---------------|--------------|-------|---------|-------------------------|
|                                                   | N (%)         | N (%)        | N (%) |         |                         |
| 9 or less                                         | 1,083         | 7,618        | 8,701 | 1.4     | 0.106                   |
| 10                                                | 45766         | 551317       | 597083| 97.3    |                         |
| 11 or more                                        | 1028          | 6301         | 7329  | 1.2     |                         |
| Missing                                           | 57            | 562          | 619   | 0.1     |                         |
| Total                                             | 47,934        | 565,798      | 613,732| 100.0  |                         |
|                                             | No link to HES | Linked to HES | Total | P-value | Standardized difference |
|---------------------------------------------|----------------|--------------|-------|---------|-------------------------|
| **Persistent Absence Y10**                  |                |              |       |         |                         |
| No                                          | 23695          | 416921       | 440616| <0.001  | 0.277                   |
| Yes                                         | 3477           | 127053       | 130530| 21.3    |                         |
| Missing                                     | 20,762         | 21,824       | 42,586| 6.9     |                         |
| **Total**                                   | 47,934         | 565,798      | 613,732| 100.0   |                         |
| **Persistent Absence Y11**                  |                |              |       |         |                         |
| No                                          | 21914          | 395806       | 417720| <0.001  | 0.255                   |
| Yes                                         | 4247           | 143616       | 147863| 24.1    |                         |
| Missing                                     | 21773          | 26376        | 48149 | 7.8     |                         |
| **Total**                                   | 47,934         | 565,798      | 613,732| 100.0   |                         |
| **SEN: Special Education Need**             |                |              |       |         |                         |
| No                                          | 40222          | 399408       | 439630| <0.001  | 0.323                   |
| Yes                                         | 7655           | 165828       | 173483| 28.3    |                         |
| Missing                                     | 57             | 562          | 619   | 0.1     |                         |
| **Total**                                   | 47,934         | 565,798      | 613,732| 100.0   |                         |
| **AAP/S: School Action or Early Years Action, School Action Plus or Early Years Action and SEN support** |                |              |       |         |                         |
| No                                          | 41471          | 421231       | 462702| <0.001  | 0.310                   |
| Yes                                         | 6406           | 144005       | 150411| 24.5    |                         |
| Missing                                     | 57             | 562          | 619   | 0.1     |                         |
| **Total**                                   | 47934          | 565,798      | 613,732| 100.0   |                         |
| **S/EHCP: Statement and Education, health and care plan** |                |              |       |         |                         |
| No                                          | 46484          | 538775       | 585259| <0.001  | 0.093                   |
| Yes                                         | 1393           | 26461        | 27854 | 4.5     |                         |
| Missing                                     | 57             | 562          | 619   | 0.1     |                         |
| **Total**                                   | 47934          | 565,798      | 613,732| 100.0   |                         |

Notes: NPD = national pupil dataset; HES = hospital episode statistics; IDACI = income deprivation affecting children index.
Supplementary Table 9.2: Distribution of demographic variables in national pupil dataset by linking status, N = 571,918 pairs (national pupil dataset to hospital episode statistics). Cohort 1996/97

| Region 2011 pupil’s residence (first recorded) NPD | No link to HES | Linked to HES | Total | P-value | Standard difference |
|--------------------------------------------------|---------------|---------------|-------|---------|---------------------|
| London                                           | 6,243 (17.7%) | 71,652 (13.4%) | 77,895 (13.6%) | <0.001 | 0.247 |
| South East                                       | 5,961 (16.9%) | 75,452 (14.1%) | 81,413 (14.2%) |         |         |
| South West                                       | 3,021 (8.6%)  | 50,302 (9.4%)  | 53,323 (9.3%)  |         |         |
| West Midlands                                    | 3,392 (9.6%)  | 60,027 (11.2%) | 63,419 (11.1%) |         |         |
| North West                                       | 3,630 (10.3%) | 77,805 (14.5%) | 81,435 (14.2%) |         |         |
| North East                                       | 1,025 (2.9%)  | 27,374 (5.1%)  | 28,399 (5.0%)  |         |         |
| Yorkshire and The Humber                         | 2,908 (8.2%)  | 54,564 (10.2%) | 57,472 (10.0%) |         |         |
| East Midlands                                    | 2,769 (7.8%)  | 42,187 (7.9%)  | 44,956 (7.9%)  |         |         |
| East of England                                  | 4,525 (12.8%) | 54,424 (10.1%) | 58,949 (10.3%) |         |         |
| Wales                                            | 1,025 (2.9%)  | 27,374 (5.1%)  | 28,399 (5.0%)  |         |         |
| Missing                                          | 1,818 (5.2%)  | 22,794 (4.2%)  | 24,612 (4.3%)  |         |         |
| Total                                            | 35,292 (100.0%) | 536,626 (100.0%) | 571,918 (100.0%) |         |         |

| Ethnic group (NPD)                               | No link to HES | Linked to HES | Total | P-value | Standard difference |
|--------------------------------------------------|---------------|---------------|-------|---------|---------------------|
| White                                            | 24,452 (69.3%)| 453,764 (84.6%)| 478,216 (83.6%) | <0.001 | 0.159 |
| Asian                                            | 2,584 (7.3%)  | 37,654 (7.0%)  | 40,238 (7.0%)  |         |         |
| Black                                            | 1,429 (4.0%)  | 19,228 (3.6%)  | 20,657 (3.6%)  |         |         |
| Chinese                                          | 213 (0.6%)    | 1,439 (0.3%)   | 1,652 (0.3%)   |         |         |
| Any other ethnic group                           | 626 (1.8%)    | 3,951 (0.7%)   | 4,577 (0.8%)   |         |         |
| Mixed                                            | 1,278 (3.6%)  | 19,286 (3.6%)  | 20,564 (3.6%)  |         |         |
| Missing                                          | 4,717 (13.4%) | 1,297 (0.2%)   | 6,014 (1.1%)   |         |         |
| Total                                            | 35,292 (100.0%) | 536,619 (100.0%) | 571,918 (100.0%) |         |         |

| Sex (NPD)                                        | No link to HES | Linked to HES | Total | P-value | Standard difference |
|--------------------------------------------------|---------------|---------------|-------|---------|---------------------|
| Male                                             | 17,014 (48.2%)| 275,479 (51.3%)| 292,493 (51.1%) | <0.001 | 0.062 |
| Female                                           | 18,268 (51.8%)| 261,094 (48.7%)| 279,362 (48.8%) |         |         |
| Missing                                          | 17 (0.0%)     | 46 (0.0%)     | 63 (0.0%)      |         |         |
| Total                                            | 35,292 (100.0%) | 536,619 (100.0%) | 571,918 (100.0%) |         |         |

| IDACI Deciles (first Census) NPD                  | No link to HES | Linked to HES | Total | P-value | Standard difference |
|--------------------------------------------------|---------------|---------------|-------|---------|---------------------|
| 1 (deprived)                                     | 4,866 (13.8%) | 50,540 (9.4%) | 55,406 (9.7%) | <0.001 | 0.218 |
| 2                                                | 4,247 (12.0%) | 51,132 (9.5%) | 55,379 (9.7%) |         |         |
| 3                                                | 3,811 (10.8%) | 51,662 (9.6%) | 55,473 (9.7%) |         |         |
| 4                                                | 3,738 (10.6%) | 51,725 (9.6%) | 55,463 (9.7%) |         |         |
| 5                                                | 3,444 (9.8%)  | 52,336 (9.8%) | 55,780 (9.8%) |         |         |
| 6                                                | 3,310 (9.4%)  | 52,503 (9.8%) | 55,813 (9.8%) |         |         |
| 7                                                | 2,936 (8.3%)  | 53,336 (9.9%) | 56,272 (9.8%) |         |         |
| 8                                                | 2,914 (8.3%)  | 54,281 (10.1%)| 57,195 (10.0%)|         |         |
| 9                                                | 2,851 (8.1%)  | 55,791 (10.4%)| 58,642 (10.3%)|         |         |
| 10 (affluent)                                    | 2,701 (7.7%)  | 56,355 (10.5%)| 59,056 (10.3%)|         |         |
| Missing                                          | 481 (1.4%)    | 6,958 (1.3%)  | 7,439 (1.3%)  |         |         |
| Total                                            | 35,292 (100.0%) | 536,619 (100.0%) | 571,918 (100.0%) |         |         |

| Age at start academic year (first recorded) NPD | No link to HES | Linked to HES | Total | P-value | Standard difference |
|------------------------------------------------|---------------|---------------|-------|---------|---------------------|
| 4 or less                                       | 406 (1.2%)    | 4,281 (0.8%)  | 4,687 (0.8%)  | <0.001 | 0.149 |
| 5                                                | 32,406 (91.8%)| 512,223 (95.5%)| 544,629 (95.2%)|         |         |
| 6 or more                                       | 2,470 (7.0%)  | 20,069 (3.7%)  | 22,539 (3.9%)  |         |         |
| Missing                                          | 17 (0.0%)     | 46 (0.0%)     | 63 (0.0%)      |         |         |
| Total                                            | 35,292 (100.0%) | 536,619 (100.0%) | 571,918 (100.0%) |         |         |

| Persistent Absence Y5                            | No link to HES | Linked to HES | Total | P-value | Standard difference |
|--------------------------------------------------|---------------|---------------|-------|---------|---------------------|
| No                                               | 21,882 (62.0%)| 466,085 (86.9%)| 487,967 (85.3%) | <0.001 | 0.174 |
| Yes                                              | 1,381 (3.9%)  | 55,794 (10.4%)| 57,175 (10.0%) |         |         |
| Missing                                          | 12,036 (34.1%)| 14,740 (2.7%)  | 26,776 (4.7%)  |         |         |
| Total                                            | 35,292 (100.0%) | 536,619 (100.0%) | 571,918 (100.0%) |         |         |
Supplementary Table 9.2: Continued

| Persistent Absence | No link to HES | Linked to HES | Total | P-value | Standard difference |
|--------------------|----------------|--------------|-------|---------|---------------------|
| Y6                 | n (%)          | n (%)        | n (%) |         |                     |
| No                 | 21,214         | 466,146      | 487,360 | 85.2    | <0.001              | 0.178               |
| Yes                | 1,184          | 51,790       | 52,974 | 9.3     |                     |                     |
| Missing            | 12,901         | 18,683       | 31,584 | 5.5     |                     |                     |
| Total              | 35,299         | 536,619      | 571,918 | 100.0   |                     |                     |

| Persistent Absence | No link to HES | Linked to HES | Total | P-value | Standard difference |
|--------------------|----------------|--------------|-------|---------|---------------------|
| Y7                 | n (%)          | n (%)        | n (%) |         |                     |
| No                 | 20,027         | 439,725      | 459,752 | 80.4    | <0.001              | 0.243               |
| Yes                | 1,309          | 67,155       | 68,464 | 12.0    |                     |                     |
| Missing            | 13,963         | 29,739       | 43,702 | 7.6     |                     |                     |
| Total              | 35,299         | 536,619      | 571,918 | 100.0   |                     |                     |

| Persistent Absence | No link to HES | Linked to HES | Total | P-value | Standard difference |
|--------------------|----------------|--------------|-------|---------|---------------------|
| Y8                 | n (%)          | n (%)        | n (%) |         |                     |
| No                 | 19,656         | 431,203      | 450,859 | 78.8    | <0.001              | 0.260               |
| Yes                | 1,420          | 73,341       | 74,761 | 13.2    |                     |                     |
| Missing            | 14,223         | 31,075       | 45,298 | 7.9     |                     |                     |
| Total              | 35,299         | 536,619      | 571,918 | 100.0   |                     |                     |

| Persistent Absence | No link to HES | Linked to HES | Total | P-value | Standard difference |
|--------------------|----------------|--------------|-------|---------|---------------------|
| Y9                 | n (%)          | n (%)        | n (%) |         |                     |
| No                 | 19,644         | 430,882      | 450,526 | 78.8    | <0.001              | 0.280               |
| Yes                | 1,270          | 73,033       | 74,303 | 13.0    |                     |                     |
| Missing            | 14,385         | 31,075       | 45,298 | 7.9     |                     |                     |
| Total              | 35,299         | 536,619      | 571,918 | 100.0   |                     |                     |

| Persistent Absence | No link to HES | Linked to HES | Total | P-value | Standard difference |
|--------------------|----------------|--------------|-------|---------|---------------------|
| Y10                | n (%)          | n (%)        | n (%) |         |                     |
| No                 | 19,880         | 440,437      | 460,317 | 80.5    | <0.001              | 0.269               |
| Yes                | 971            | 60,133       | 61,104 | 10.7    |                     |                     |
| Missing            | 14,448         | 36,049       | 50,497 | 8.8     |                     |                     |
| Total              | 35,299         | 536,619      | 571,918 | 100.0   |                     |                     |

| Persistent Absence | No link to HES | Linked to HES | Total | P-value | Standard difference |
|--------------------|----------------|--------------|-------|---------|---------------------|
| Y11                | n (%)          | n (%)        | n (%) |         |                     |
| No                 | 17,968         | 414,077      | 432,045 | 75.5    | <0.001              | 0.264               |
| Yes                | 1,053          | 62,465       | 63,518 | 11.1    |                     |                     |
| Missing            | 16,278         | 60,077       | 76,355 | 13.4    |                     |                     |
| Total              | 35,299         | 536,619      | 571,918 | 100.0   |                     |                     |

| SEN: Special Education Need | No link to HES | Linked to HES | Total | P-value | Standard difference |
|-----------------------------|----------------|--------------|-------|---------|---------------------|
| No                          | n (%)          | n (%)        | n (%) |         |                     |
| No                          | 27,164         | 306,131      | 333,295 | 58.3    | <0.001              | 0.435               |
| Yes                         | 8,118          | 230,442      | 238,560 | 41.7    |                     |                     |
| Missing                     | 17             | 46           | 63     | 0.0     |                     |                     |
| Total                       | 35,299         | 536,619      | 571,918 | 100.0   |                     |                     |

| AAP/S: School Action or Early Years Action, School Action Plus or Early Years Action and SEN support | No link to HES | Linked to HES | Total | P-value | Standard difference |
|--------------------------------------------------------------------------------------------------|----------------|--------------|-------|---------|---------------------|
| No                                                                                              | n (%)          | n (%)        | n (%) |         |                     |
| No                                                                                              | 27,448         | 315,908      | 343,356 | 60.0    | <0.001              | 0.416               |
| Yes                                                                                             | 7,834          | 220,665      | 228,499 | 40.0    |                     |                     |
| Missing                                                                                         | 17             | 46           | 63     | 0.0     |                     |                     |
| Total                                                                                           | 35,299         | 536,619      | 571,918 | 100.0   |                     |                     |

| S/EHCP: Statement and Education, health and care plan | No link to HES | Linked to HES | Total | P-value | Standard difference |
|------------------------------------------------------|----------------|--------------|-------|---------|---------------------|
| No                                                                                              | n (%)          | n (%)        | n (%) |         |                     |
| No                                                                                              | 34,810         | 512,580      | 547,390 | 95.7    | <0.001              | 0.188               |
| Yes                                                                                             | 472            | 23,993       | 24,465  | 4.3     |                     |                     |
| Missing                                                                                         | 17             | 46           | 63     | 0.0     |                     |                     |
| Total                                                                                           | 35,299         | 536,619      | 571,918 | 100.0   |                     |                     |

Notes: NPD = national pupil dataset; HES = hospital episode statistics; IDACI = income deprivation affecting children index.
### Supplementary Table 9.3: Distribution of demographic variables in national pupil dataset by linking status, N = 529,910 pairs (national pupil dataset to hospital episode statistics). Cohort 1999/00

| Region 2011 pupil’s residence (first recorded) NPD | No linked to HES | Linked to HES | Total | P-value | Standard difference |
|---------------------------------------------------|------------------|--------------|-------|---------|---------------------|
|                                                   | n (%)            | n (%)        | n (%) |         |                     |
| London                                            | 4,303 19.4       | 71,001 14.0  | 75,304 14.2  | <0.001 | 0.3105              |
| South East                                        | 3,881 17.5       | 74,189 14.6  | 78,070 14.7  |         |                     |
| South West                                        | 1,364 6.1        | 45,672 9.0   | 47,036 8.9   |         |                     |
| West Midlands                                     | 2,274 10.3       | 55,174 10.9  | 57,448 10.8  |         |                     |
| North West                                        | 2,036 9.2        | 70,533 13.9  | 72,569 13.7  |         |                     |
| North East                                        | 585 2.6          | 24,497 4.8   | 25,082 4.7   |         |                     |
| Yorkshire and The Humber                          | 1,502 6.8        | 49,701 9.8   | 51,203 9.7   |         |                     |
| East Midlands                                     | 1,786 8.1        | 40,944 8.1   | 42,730 8.1   |         |                     |
| East of England                                   | 3,119 14.1       | 52,238 10.3  | 55,357 10.4  |         |                     |
| Wales                                             | 0 0.0            | 64 0.0       | 64 0.0       |         |                     |
| Missing                                           | 1,327 6.0        | 23,720 4.7   | 25,047 4.7   |         |                     |
| Total                                             | 22,177 100.0     | 507,733 100.0| 529,910 100.0|         |                     |

| Ethnic group (NPD)                                | No linked to HES | Linked to HES | Total | P-value | Standard difference |
|---------------------------------------------------|------------------|--------------|-------|---------|---------------------|
|                                                   | n (%)            | n (%)        | n (%) |         |                     |
| White                                             | 15,692 70.7      | 415,660 81.9 | 431,352 81.4  | <0.001 | 0.2809              |
| Asian                                             | 2,581 11.6       | 43,061 8.5   | 45,642 8.6   |         |                     |
| Black                                             | 1,735 7.8        | 21,528 4.2   | 23,263 4.4   |         |                     |
| Chinese                                           | 172 0.8          | 1,530 0.3    | 1,702 0.3    |         |                     |
| Any other ethnic group                            | 700 3.2          | 5,146 1.0    | 5,846 1.1    |         |                     |
| Mixed                                             | 1,178 5.3        | 20,177 4.0   | 21,355 4.0   |         |                     |
| Missing                                           | 127 0.6          | 623 0.1      | 750 0.1      |         |                     |
| Total                                             | 22,185 100.0     | 507,725 100.0| 529,910 100.0|         |                     |

| Sex (NPD)                                         | No linked to HES | Linked to HES | Total | P-value | Standard difference |
|---------------------------------------------------|------------------|--------------|-------|---------|---------------------|
|                                                   | n (%)            | n (%)        | n (%) |         |                     |
| Male                                              | 9,717 43.8       | 261,398 51.5 | 271,115 51.2  | <0.001 | 0.1526              |
| Female                                            | 12,445 56.1      | 246,116 48.5 | 258,561 48.8  |         |                     |
| Missing                                           | 23 0.1           | 211 0.0      | 234 0.0      |         |                     |
| Total                                             | 22,185 100.0     | 507,725 100.0| 529,910 100.0|         |                     |

| IDACI Deciles (first Census) NPD                   | No linked to HES | Linked to HES | Total | P-value | Standard difference |
|---------------------------------------------------|------------------|--------------|-------|---------|---------------------|
|                                                   | n (%)            | n (%)        | n (%) |         |                     |
| 1 (deprived)                                      | 2,863 12.9       | 49,733 9.8   | 52,596 9.9   | <0.001 | 0.1417              |
| 2                                                 | 2,487 11.2       | 49,457 9.7   | 51,944 9.8   |         |                     |
| 3                                                 | 2,257 10.2       | 49,130 9.7   | 51,387 9.7   |         |                     |
| 4                                                 | 2,263 10.2       | 49,153 9.7   | 51,416 9.7   |         |                     |
| 5                                                 | 2,139 9.6        | 49,450 9.7   | 51,589 9.7   |         |                     |
| 6                                                 | 2,056 9.3        | 49,965 9.8   | 52,021 9.8   |         |                     |
| 7                                                 | 1,980 8.9        | 50,467 9.9   | 52,447 9.9   |         |                     |
| 8                                                 | 2,077 9.4        | 51,321 10.1  | 53,398 10.1  |         |                     |
| 9                                                 | 1,972 8.9        | 52,884 10.4  | 54,856 10.4  |         |                     |
| 10 (affluent)                                     | 1,953 8.8        | 53,904 10.6  | 55,857 10.5  |         |                     |
| Missing                                           | 138 0.6          | 2,261 0.4    | 2,399 0.5    |         |                     |
| Total                                             | 22,185 100.0     | 507,725 100.0| 529,910 100.0|         |                     |

| Age at start academic year (first recorded) NPD    | No linked to HES | Linked to HES | Total | P-value | Standard difference |
|---------------------------------------------------|------------------|--------------|-------|---------|---------------------|
|                                                   | n (%)            | n (%)        | n (%) |         |                     |
| 4 or less                                         | 152 0.7          | 2,409 0.5    | 2,561 0.5  | <0.001 | 0.0674              |
| 5                                                 | 21,775 98.2      | 502,514 99.0 | 524,289 98.9 |         |                     |
| 6 or more                                         | 249 1.1          | 2,775 0.5    | 3,024 0.6   |         |                     |
| Missing                                           | 9 0.0            | 27 0.0       | 36 0.0      |         |                     |
| Total                                             | 22,185 100.0     | 507,725 100.0| 529,910 100.0|         |                     |
| Persistent Absence | Linked to HES | Total |
|--------------------|-------------|-------|
|                     | n (%)       | n (%) |
| Persistent Absence Y2 |             |       |
| No                  | 17,347 78.2 | 445,430 87.7 | 462,777 87.3 |
| Yes                 | 1,440 6.5   | 53,770 10.6  | 55,210 10.4   |
| Missing             | 3398 15.3   | 8525 1.7    | 11923 2.3     |
| Total               | 22,185 100.0 | 507725 100.0 | 529910 100.0  |

P-value Standard difference

<0.001 0.1075

Persistent Absence Y3

No 16513 74.4 441778 86.9 456916 86.2 <0.001 0.1645
Yes 910 4.1 48320 9.1 49230 9.3
Missing 5764 26.5 20333 4.0 25097 4.7
Total 22185 100.0 507725 100.0 529910 100.0

Persistent Absence Y4

No 15952 71.9 441778 86.9 456916 86.2 <0.001 0.1885
Yes 639 2.9 47234 9.3 47873 9.0
Missing 6081 27.2 34106 6.8 39211 6.7
Total 22185 100.0 507725 100.0 529910 100.0

Persistent Absence Y5

No 15034 67.8 436743 86.0 451777 85.3 <0.001 0.2105
Yes 512 2.3 38455 7.6 39967 7.4
Missing 6081 27.9 34106 6.7 39211 6.7
Total 22185 100.0 507725 100.0 529910 100.0

Persistent Absence Y6

No 14782 66.6 426385 84.0 441167 83.3 <0.001 0.2316
Yes 639 2.9 47234 9.3 47873 9.0
Missing 6081 27.5 34106 6.7 39211 6.7
Total 22185 100.0 507725 100.0 529910 100.0

Persistent Absence Y7

No 14917 67.2 432105 85.1 447022 84.4 <0.001 0.2515
Yes 436 2.0 40401 8.0 40837 7.7
Missing 6081 27.8 34106 6.7 39211 6.7
Total 22185 100.0 507725 100.0 529910 100.0

Persistent Absence Y8

No 14,799 66.7 424,866 83.7 439,665 83.0 <0.001 0.2715
Yes 488 2.2 45,726 9.0 46,214 8.7
Missing 6081 31.1 37,133 7.3 44,031 8.3
Total 22185 100.0 507725 100.0 529910 100.0

Supplementary Table 9.3: Continued
### Supplementary Table 9.3: Continued

|                          | No linked to HES | Linked to HES | Total | P-value | Standard difference |
|--------------------------|------------------|---------------|-------|---------|--------------------|
| **Persistent Absence Y11** |                  |               |       |         |                    |
| No                      | 14,475 (65.2)    | 416,671 (82.1)| 431,146 (81.4)| <0.001  | 0.2353             |
| Yes                     | 621 (2.8)        | 46,459 (9.2)  | 47,080 (8.9)  |         |                    |
| Missing                 | 7,089 (32.0)     | 44,595 (8.8)  | 51,684 (9.8)  |         |                    |
| Total                   | 22,185 (100.0)   | 507,725 (100.0)| 529,910 (100.0)|         |                    |

| **SEN:** Special Education Need |                  |               |       |         |                    |
| No                         | 16,438 (74.1)    | 291,701 (57.5)| 308,139 (58.1)| <0.001  | 0.3612             |
| Yes                        | 5,739 (25.9)     | 215,997 (42.5)| 221,736 (41.8)|         |                    |
| Missing                    | 0 (0.0)          | 35 (0.0)      | 35 (0.0)     |         |                    |
| Total                      | 22,177 (100.0)   | 507,733 (100.0)| 529,910 (100.0)|         |                    |

| **AAP/S:** School Action or Early Years Action, School Action Plus or Early Years Action and SEN support |                  |               |       |         |                    |
| No                        | 16,495 (74.4)    | 298,569 (58.8)| 315,064 (59.5)| <0.001  | 0.3390             |
| Yes                       | 5,682 (25.6)     | 209,129 (41.2)| 214,811 (40.5)|         |                    |
| Missing                   | 0 (0.0)          | 35 (0.0)      | 35 (0.0)     |         |                    |
| Total                     | 22,177 (100.0)   | 507,733 (100.0)| 529,910 (100.0)|         |                    |

| **S/EHCP:** Statement and Education, health and care plan |                  |               |       |         |                    |
| No                        | 21,980 (99.1)    | 484,288 (95.4)| 506,268 (95.5)| <0.001  | 0.2324             |
| Yes                       | 197 (0.9)        | 23,410 (4.6)  | 23,607 (4.5)  |         |                    |
| Missing                   | 0 (0.0)          | 35 (0.0)      | 35 (0.0)     |         |                    |
| Total                     | 22,177 (100.0)   | 507,733 (100.0)| 529,910 (100.0)|         |                    |

Notes: NPD = national pupil dataset; HES = hospital episode statistics; IDACI = income deprivation affecting children index.
Supplementary Table 9.4: Distribution of demographic variables in national pupil dataset by linking status, N = 578,809 pairs (national pupil dataset to hospital episode statistics). Cohort 2004/05

| Region 2011 pupil’s residence (first recorded) NPD | No link to HES | Linked to HES | Total  | P-value | Standard difference |
|---------------------------------------------------|----------------|--------------|--------|---------|---------------------|
|                                                   | n (%)          | n (%)        | n (%)  |         |                     |
| London                                            | 1,590 18.8     | 83,817 14.7  | 85,407 14.8  | 0.000  | 0.237               |
| South East                                        | 1,353 16.0     | 83,748 14.7  | 85,101 14.7  |         |                     |
| South West                                        | 504 5.9        | 49,993 8.8   | 50,497 8.7   |         |                     |
| West Midlands                                     | 759 9.0        | 60,358 10.6  | 61,117 10.6  |         |                     |
| North West                                        | 986 11.6       | 76,373 13.4  | 77,359 13.4  |         |                     |
| North East                                        | 197 2.3        | 26,007 4.6   | 26,204 4.5   |         |                     |
| Yorkshire and The Humber                          | 671 7.9        | 56,330 9.9   | 57,001 9.8   |         |                     |
| East Midlands                                     | 689 8.1        | 45,255 7.9   | 45,944 7.9   |         |                     |
| East of England                                   | 1,040 12.3     | 57,545 10.1  | 58,585 10.1  |         |                     |
| Wales                                             | 0 0.0          | 69 0.0       | 69 0.0      |         |                     |
| Missing                                           | 685 8.1        | 30,840 5.4   | 31,525 5.4   |         |                     |
| Total                                             | 8,474 100.0    | 570,335 100.0| 578,809 100.0|         |                     |

| Ethnic group (NPD)                                | No link to HES | Linked to HES | Total  | P-value | Standard difference |
|---------------------------------------------------|----------------|--------------|--------|---------|---------------------|
|                                                   | n (%)          | n (%)        | n (%)  |         |                     |
| White                                             | 5,255 62.0     | 439,397 77.0 | 444,652 76.8 | 0.000  | 0.358               |
| Asian                                             | 1,207 14.2     | 57,790 10.1  | 58,997 10.2  |         |                     |
| Black                                             | 696 8.2        | 31,656 5.6   | 32,352 5.6   |         |                     |
| Chinese                                           | 89 1.0         | 2,038 0.4    | 2,127 0.4   |         |                     |
| Any other ethnic group                            | 486 5.7        | 8,375 1.5    | 8,861 1.5   |         |                     |
| Mixed                                             | 575 6.8        | 29,871 5.2   | 30,446 5.3   |         |                     |
| Missing                                           | 169 2.0        | 1,205 0.2    | 1,374 0.2   |         |                     |
| Total                                             | 8,477 100.0    | 570,332 100.0| 578,809 100.0|         |                     |

| Sex (NPD)                                         | No link to HES | Linked to HES | Total  | P-value | Standard difference |
|---------------------------------------------------|----------------|--------------|--------|---------|---------------------|
|                                                   | n (%)          | n (%)        | n (%)  |         |                     |
| Male                                              | 3,660 43.2     | 292,784 51.3 | 296,444 51.2 | 0.000  | 0.166               |
| Female                                            | 4,814 56.8     | 277,508 48.7 | 282,322 48.8 |         |                     |
| Missing                                           | 0 0.0          | 43 0.0       | 43 0.0   |         |                     |
| Total                                             | 8,474 100.0    | 570,335 100.0| 578,809 100.0|         |                     |

| IDACI Deciles (first Census) NPD                   | No link to HES | Linked to HES | Total  | P-value | Standard difference |
|---------------------------------------------------|----------------|--------------|--------|---------|---------------------|
|                                                   | n (%)          | n (%)        | n (%)  |         |                     |
| 1 (deprived)                                      | 909 10.7       | 53,590 9.4   | 54,499 9.4   | 0.000  | 0.070               |
| 2                                                 | 855 10.1       | 53,748 9.4   | 54,603 9.4   |         |                     |
| 3                                                 | 849 10.0       | 54,246 9.5   | 55,095 9.5   |         |                     |
| 4                                                 | 750 8.8        | 54,250 9.5   | 55,000 9.5   |         |                     |
| 5                                                 | 812 9.6        | 55,571 9.7   | 56,383 9.7   |         |                     |
| 6                                                 | 840 9.9        | 56,601 9.9   | 57,441 9.9   |         |                     |
| 7                                                 | 844 10.0       | 57,776 10.1  | 58,620 10.1  |         |                     |
| 8                                                 | 885 10.4       | 58,854 10.3  | 59,739 10.3  |         |                     |
| 9                                                 | 858 10.1       | 61,048 10.7  | 61,906 10.7  |         |                     |
| 10 (affluent)                                     | 821 9.7        | 62,514 11.0  | 63,335 10.9  |         |                     |
| Missing                                           | 54 0.6         | 2,134 0.4    | 2,188 0.4   |         |                     |
| Total                                             | 8,477 100.0    | 570,332 100.0| 578,809 100.0|         |                     |

| Age at start academic year (first recorded) NPD    | No link to HES | Linked to HES | Total  | P-value | Standard difference |
|---------------------------------------------------|----------------|--------------|--------|---------|---------------------|
|                                                   | n (%)          | n (%)        | n (%)  |         |                     |
| 4 or less                                         | 34 0.4         | 1,078 0.2    | 1,112 0.2   | 0.000  | 0.083               |
| 5                                                 | 8,382 98.9     | 568,198 99.6 | 576,580 99.6 |         |                     |
| 6 or more                                         | 61 0.7         | 1,042 0.2    | 1,103 0.2   |         |                     |
| Missing                                           | 0 0.0          | 14 0.0       | 14 0.0    |         |                     |
| Total                                             | 8,477 100.0    | 570,332 100.0| 578,809 100.0|         |                     |
| Persistent Absence Y1 | No link to HES | Linked to HES | Total | P-value | Standard difference |
|----------------------|----------------|--------------|--------|---------|---------------------|
|                      | n | (%)      | n | (%)     | n | (%)      |             |          |         |
| No                   | 6,560 | 77.4 | 499,976 | 87.7 | 506,536 | 87.5 | 0.000 | 0.034 |
| Yes                  | 766 | 9.0 | 64,769 | 11.4 | 65,535 | 11.3 |
| Missing              | 1151 | 13.6 | 5587 | 1.0 | 6738 | 1.2 |
| Total                | 8477 | 100.0 | 570332 | 100.0 | 578809 | 100.0 |

| Persistent Absence Y2 | No link to HES | Linked to HES | Total | P-value | Standard difference |
|----------------------|----------------|--------------|--------|---------|---------------------|
|                      | n | (%)      | n | (%)     | n | (%)      |             |          |         |
| No                   | 6,288 | 74.2 | 524,842 | 92.0 | 531,130 | 91.8 | 0.000 | 0.074 |
| Yes                  | 314 | 3.7 | 35,861 | 6.3 | 36,175 | 6.2 |
| Missing              | 1875 | 22.1 | 15,547 | 2.7 | 17,843 | 3.1 |
| Total                | 8477 | 100.0 | 570332 | 100.0 | 578809 | 100.0 |

| Persistent Absence Y3 | No link to HES | Linked to HES | Total | P-value | Standard difference |
|----------------------|----------------|--------------|--------|---------|---------------------|
|                      | n | (%)      | n | (%)     | n | (%)      |             |          |         |
| No                   | 5904 | 69.6 | 520032 | 91.2 | 525936 | 90.9 | 0.000 | 0.079 |
| Yes                  | 77 | 0.9 | 933 | 1.6 | 950 | 1.6 |
| Missing              | 2575 | 30.5 | 14,561 | 2.6 | 16,122 | 2.7 |
| Total                | 8477 | 100.0 | 570332 | 100.0 | 578809 | 100.0 |

| Persistent Absence Y4 | No link to HES | Linked to HES | Total | P-value | Standard difference |
|----------------------|----------------|--------------|--------|---------|---------------------|
|                      | n | (%)      | n | (%)     | n | (%)      |             |          |         |
| No                   | 5804 | 68.5 | 529210 | 92.8 | 535014 | 92.4 | 0.000 | 0.125 |
| Yes                  | 113 | 1.3 | 21646 | 3.8 | 21759 | 3.8 |
| Missing              | 2560 | 30.2 | 19476 | 3.4 | 22036 | 3.8 |
| Total                | 8477 | 100.0 | 570332 | 100.0 | 578809 | 100.0 |

| Persistent Absence Y5 | No link to HES | Linked to HES | Total | P-value | Standard difference |
|----------------------|----------------|--------------|--------|---------|---------------------|
|                      | n | (%)      | n | (%)     | n | (%)      |             |          |         |
| No                   | 5,666 | 66.8 | 521,853 | 91.5 | 527,519 | 91.1 | 0.000 | 0.171 |
| Yes                  | 95 | 1.1 | 25,261 | 4.4 | 25,356 | 4.4 |
| Missing              | 2,716 | 32.0 | 23,218 | 4.1 | 25,934 | 4.5 |
| Total                | 8,477 | 100.0 | 570332 | 100.0 | 578809 | 100.0 |

| Persistent Absence Y6 | No link to HES | Linked to HES | Total | P-value | Standard difference |
|----------------------|----------------|--------------|--------|---------|---------------------|
|                      | n | (%)      | n | (%)     | n | (%)      |             |          |         |
| No                   | 5,575 | 65.8 | 521,192 | 91.4 | 526,767 | 91.0 | 0.000 | 0.143 |
| Yes                  | 97 | 1.1 | 21,989 | 3.9 | 22,086 | 3.8 |
| Missing              | 2,805 | 33.1 | 27,151 | 4.8 | 29,956 | 5.2 |
| Total                | 8,477 | 100.0 | 570332 | 100.0 | 578809 | 100.0 |

| Persistent Absence Y7 | No link to HES | Linked to HES | Total | P-value | Standard difference |
|----------------------|----------------|--------------|--------|---------|---------------------|
|                      | n | (%)      | n | (%)     | n | (%)      |             |          |         |
| No                   | 5,407 | 63.8 | 505,729 | 88.7 | 511,136 | 88.3 | 0.000 | 0.188 |
| Yes                  | 99 | 1.2 | 27,407 | 4.8 | 27,506 | 4.8 |
| Missing              | 2,971 | 35.0 | 37,196 | 6.5 | 40,167 | 6.9 |
| Total                | 8,477 | 100.0 | 570332 | 100.0 | 578809 | 100.0 |

| Persistent Absence Y8 | No link to HES | Linked to HES | Total | P-value | Standard difference |
|----------------------|----------------|--------------|--------|---------|---------------------|
|                      | n | (%)      | n | (%)     | n | (%)      |             |          |         |
| No                   | 5,341 | 63.0 | 494,149 | 86.6 | 499,490 | 86.3 | 0.000 | 0.201 |
| Yes                  | 145 | 1.7 | 36,519 | 6.4 | 36,664 | 6.3 |
| Missing              | 2,991 | 35.3 | 39,664 | 7.0 | 42,655 | 7.4 |
| Total                | 8,477 | 100.0 | 570332 | 100.0 | 578809 | 100.0 |

| SEN: Special Education Need | No link to HES | Linked to HES | Total | P-value | Standard difference |
|-----------------------------|----------------|--------------|--------|---------|---------------------|
| No                          | n | (%)      | n | (%)     | n | (%)      |             |          |         |
| No                          | 6,626 | 78.2 | 372,191 | 65.3 | 378,817 | 65.4 | 0.000 | 0.306 |
| Yes                         | 1,851 | 21.8 | 198,127 | 34.7 | 199,978 | 34.5 |
| Missing                     | 0 | 0.0 | 14 | 0.0 | 14 | 0.0 |
| Total                       | 8,477 | 100.0 | 570332 | 100.0 | 578809 | 100.0 |
### Supplementary Table 9.4: Continued

| AAP/S: School Action or Early Years Action, School Action Plus or Early Years Action and SEN support | No link to HES | Linked to HES | Total | P-value | Standard difference |
|---|---|---|---|---|---|
| No | 6,643 | 78.4 | 379,274 | 66.5 | 385,917 | 66.7 | 0.000 | 0.284 |
| Yes | 1,834 | 21.6 | 191,044 | 33.5 | 192,878 | 33.3 | |
| Missing | 0 | 0.0 | 14 | 0.0 | 14 | 0.0 | |
| Total | 8,477 | 100.0 | 570,332 | 100.0 | 578,809 | 100.0 | |

| S/EHCP: Statement and Education, health and care plan | No link to HES | Linked to HES | Total | P-value | Standard difference |
|---|---|---|---|---|---|
| No | 8,412 | 99.2 | 548,057 | 96.1 | 556,469 | 96.1 | 0.000 | 0.218 |
| Yes | 65 | 0.8 | 22,261 | 3.9 | 22,326 | 3.9 | |
| Missing | 0 | 0.0 | 14 | 0.0 | 14 | 0.0 | |
| Total | 8,477 | 100.0 | 570,332 | 100.0 | 578,809 | 100.0 | |

Notes: NPD = national pupil dataset; HES = hospital episode statistics; IDACI = income deprivation affecting children index.
## Supplementary Table 10: Odd Ratios (OR) and adjusted Odd Ratios (aOR) for linkage to HES by cohort

| Ethnic group                      | (1) Bivariate | (2) Multivariable | (3) Bivariate | (4) Multivariable |
|-----------------------------------|---------------|-------------------|---------------|-------------------|
|                                   | OR Conf. Int. | OR Conf. Int.     | OR Conf. Int. | OR Conf. Int.     |
| White                             | Ref           | Ref               | Ref           | Ref               |
| Asian                             | 0.74 [0.71,0.77]** | 0.69 [0.66,0.80]** | 0.79 [0.75,0.83]** | 0.69 [0.66,0.73]** |
| Black                             | 0.64 [0.61,0.68]** | 0.62 [0.59,0.66]** | 0.73 [0.69,0.77]** | 0.67 [0.63,0.71]** |
| Chinese                           | 0.28 [0.25,0.32]** | 0.29 [0.26,0.33]** | 0.36 [0.32,0.42]** | 0.38 [0.33,0.44]** |
| Any other ethnic group            | 0.41 [0.38,0.45]** | 0.42 [0.40,0.46]** | 0.34 [0.31,0.37]** | 0.32 [0.30,0.35]** |
| Missing                           | 0.03 [0.03,0.03]** | 0.03 [0.03,0.03]** | 0.01 [0.01,0.02]** | 0.01 [0.01,0.02]** |
| Female                            | 1.30 [1.28,1.33]** | 1.35 [1.32,1.37]** | 0.88 [0.86,0.90]** | 0.87 [0.85,0.89]** |
| Missing                           | 0.94 [0.72,1.24]** | 22.77 [17.02,30.47]** | 0.17 [0.10,0.29]** | 10.21 [5.77,18.07]** |

| Region                            | (1) Bivariate | (2) Multivariable | (3) Bivariate | (4) Multivariable |
|-----------------------------------|---------------|-------------------|---------------|-------------------|
|                                   | OR Conf. Int. | OR Conf. Int.     | OR Conf. Int. | OR Conf. Int.     |
| London                            | Ref           | Ref               | Ref           | Ref               |
| South East                        | 1.16 [1.12,1.20]** | 1.31 [1.26,1.36]** | 1.10 [1.06,1.14]** | 1.12 [1.08,1.17]** |
| South West                        | 1.40 [1.35,1.46]** | 1.34 [1.28,1.40]** | 1.45 [1.39,1.52]** | 1.38 [1.31,1.45]** |
| West Midlands                     | 1.46 [1.40,1.51]** | 1.27 [1.22,1.33]** | 1.54 [1.48,1.61]** | 1.37 [1.30,1.43]** |
| North West                        | 1.53 [1.47,1.58]** | 1.36 [1.30,1.41]** | 1.87 [1.79,1.95]** | 1.64 [1.57,1.72]** |
| North East                        | 2.12 [2.01,2.25]** | 1.91 [1.80,2.04]** | 2.33 [2.18,2.49]** | 1.99 [1.85,2.14]** |
| Yorkshire and The Humber          | 1.68 [1.62,1.75]** | 1.34 [1.28,1.40]** | 1.64 [1.56,1.71]** | 1.42 [1.35,1.49]** |
| East Midlands                     | 1.51 [1.45,1.58]** | 1.28 [1.23,1.35]** | 1.33 [1.27,1.39]** | 1.22 [1.16,1.28]** |
| East of England                   | 1.22 [1.18,1.27]** | 1.14 [1.09,1.19]** | 1.05 [1.01,1.09]** | 1.00 [0.95,1.04]** |
| Wales                             | 0.15 [0.09,0.25]** | 0.31 [0.16,0.59]** | 0.47 [0.21,1.06]** | 0.40 [0.17,0.93]** |
| Missing                           | 1.17 [1.11,1.23]** | 1.16 [1.10,1.23]** | 1.09 [1.03,1.15]** | 1.08 [1.01,1.14]** |

| IDACI Deciles                     | (1) Bivariate | (2) Multivariable | (3) Bivariate | (4) Multivariable |
|-----------------------------------|---------------|-------------------|---------------|-------------------|
| 1 (deprived)                      | 0.61 [0.58,0.63]** | 0.67 [0.64,0.70]** | 0.68 [0.65,0.72]** | 0.71 [0.67,0.74]** |
| 2                                 | 0.76 [0.72,0.79]** | 0.78 [0.74,0.81]** | 0.79 [0.76,0.83]** | 0.77 [0.73,0.81]** |
| 3                                 | 0.85 [0.81,0.88]** | 0.86 [0.82,0.90]** | 0.89 [0.85,0.94]** | 0.87 [0.83,0.92]** |
| 4                                 | 0.93 [0.89,0.97]** | 0.95 [0.90,0.99]** | 0.91 [0.87,0.96]** | 0.90 [0.85,0.94]** |
| 5                                 | Ref            | Ref               | Ref           | Ref               |
| 6                                 | 1.09 [1.04,1.14]** | 1.11 [1.05,1.16]** | 1.04 [0.99,1.10]** | 1.05 [1.00,1.11]** |
| 7                                 | 1.20 [1.15,1.26]** | 1.26 [1.20,1.32]** | 1.20 [1.14,1.26]** | 1.23 [1.16,1.29]** |
| 8                                 | 1.24 [1.19,1.30]** | 1.31 [1.25,1.38]** | 1.23 [1.17,1.29]** | 1.27 [1.20,1.34]** |
| 9                                 | 1.27 [1.22,1.33]** | 1.31 [1.25,1.38]** | 1.29 [1.22,1.36]** | 1.37 [1.29,1.44]** |
| 10 (affluent)                     | 1.22 [1.16,1.27]** | 1.27 [1.21,1.34]** | 1.37 [1.30,1.45]** | 1.52 [1.44,1.61]** |
| Missing                           | 0.83 [0.77,0.90]** | 0.95 [0.86,1.04]** | 0.95 [0.86,1.05]** | 1.06 [0.95,1.18]** |

| Observations                      | 613,732       | 571,918           |
| Pseudo R-squared                  | 0.162         | 0.093             |
Supplementary Table 10: Continued

| Ethnic group                      | (5) Bivariate OR | Conf. Int. | aOR | Conf. Int. | (6) Multivariable OR | Conf. Int. | aOR | Conf. Int. | (7) Bivariate OR | Conf. Int. | aOR | Conf. Int. | (8) Multivariable OR | Conf. Int. | aOR | Conf. Int. |
|-----------------------------------|------------------|------------|-----|------------|----------------------|------------|-----|------------|------------------|------------|-----|------------|----------------------|------------|-----|------------|
| White                             | Ref              |            | Ref |            | Ref                  |            |     | Ref        |                  |            |     | Ref        |                   |            |     | Ref        |
| Asian                             | 0.63 [0.60,0.66]**| 0.56 [0.54,0.59]**| 0.57 [0.54,0.61]**| 0.51 [0.47,0.54]** |
| Black                             | 0.47 [0.44,0.49]**| 0.43 [0.40,0.45]**| 0.54 [0.50,0.59]**| 0.47 [0.43,0.51]** |
| Chinese                           | 0.34 [0.29,0.39]**| 0.35 [0.30,0.41]**| 0.27 [0.22,0.34]**| 0.27 [0.22,0.34]** |
| Any other ethnic group            | 0.28 [0.26,0.30]**| 0.26 [0.24,0.28]**| 0.21 [0.19,0.23]**| 0.18 [0.17,0.20]** |
| Mixed                             | 0.65 [0.61,0.69]**| 0.64 [0.60,0.68]**| 0.62 [0.57,0.68]**| 0.60 [0.55,0.66]** |
| Missing                           | 0.19 [0.15,0.22]**| 0.21 [0.17,0.25]**| 0.09 [0.07,0.10]**| 0.09 [0.07,0.10]** |
| Sex                               |                  |            |     |            |                      |            |     |            |                  |            |     |            |                      |            |     |            |
| Male                              | Ref              |            | Ref |            | Ref                  |            |     | Ref        |                  |            |     | Ref        |                   |            |     | Ref        |
| Female                            | 0.74 [0.72,0.76]**| 0.73 [0.71,0.75]**| 0.72 [0.69,0.75]**| 0.72 [0.69,0.75]** |
| Missing                           | 0.34 [0.22,0.52]**| 0.61 [0.39,0.90]**| 0.17 [0.05,0.54]**| 1.00 [0.29,3.50]** |
| Region                            |                  |            |     |            |                      |            |     |            |                  |            |     |            |                      |            |     |            |
| London                            | Ref              |            | Ref |            | Ref                  |            |     | Ref        |                  |            |     | Ref        |                   |            |     | Ref        |
| South East                        | 1.16 [1.11,1.21]**| 1.00 [0.95,1.04]**| 1.17 [1.09,1.26]**| 0.94 [0.87,1.02]** |
| South West                        | 2.03 [1.91,2.16]**| 1.62 [1.51,1.73]**| 1.88 [1.70,2.08]**| 1.38 [1.24,1.54]** |
| West Midlands                     | 1.47 [1.40,1.55]**| 1.23 [1.16,1.30]**| 1.51 [1.38,1.65]**| 1.21 [1.11,1.33]** |
| North West                        | 2.10 [1.99,2.22]**| 1.64 [1.55,1.74]**| 1.47 [1.36,1.59]**| 1.09 [1.00,1.19]** |
| North East                        | 2.54 [2.33,2.77]**| 1.82 [1.67,2.00]**| 2.50 [2.16,2.91]**| 1.71 [1.47,1.99]** |
| Yorkshire and The Humber          | 2.01 [1.89,2.13]**| 1.61 [1.51,1.72]**| 1.59 [1.45,1.74]**| 1.23 [1.12,1.35]** |
| East Midlands                     | 1.39 [1.31,1.47]**| 1.14 [1.08,1.21]**| 1.25 [1.14,1.36]**| 0.96 [0.87,1.06]** |
| East of England                   | 1.02 [0.97,1.06]**| 0.86 [0.81,0.90]**| 1.05 [0.97,1.14]**| 0.83 [0.76,0.90]** |
| Wales                             | 0.42 [0.20,0.89]**| 0.37 [0.17,0.80]**| 0.42 [0.13,1.33]**| 0.36 [0.11,1.19]** |
| Missing                           | 1.08 [1.02,1.15]**| 0.97 [0.91,1.03]**| 0.85 [0.78,0.93]**| 0.74 [0.68,0.82]** |
| IDACI Deciles                     |                  |            |     |            |                      |            |     |            |                  |            |     |            |                      |            |     |            |
| 1 (deprived)                      | 0.75 [0.71,0.80]**| 0.73 [0.68,0.77]**| 0.86 [0.78,0.95]**| 0.82 [0.75,0.90]** |
| 2                                 | 0.86 [0.81,0.91]**| 0.82 [0.77,0.87]**| 0.92 [0.83,1.01]**| 0.88 [0.80,0.97]** |
| 3                                 | 0.94 [0.89,1.00]**| 0.89 [0.83,0.94]**| 0.93 [0.85,1.03]**| 0.90 [0.82,1.00]** |
| 4                                 | 0.94 [0.88,1.00]**| 0.92 [0.86,0.97]**| 1.06 [0.96,1.17]**| 1.03 [0.93,1.14]** |
| 5                                 | Ref              |            | Ref |            | Ref                  |            |     | Ref        |                  |            |     | Ref        |                   |            |     | Ref        |
| 6                                 | 1.05 [0.99,1.12]**| 1.10 [1.03,1.17]**| 0.99 [0.89,1.09]**| 1.03 [0.94,1.14]** |
| 7                                 | 1.10 [1.04,1.17]**| 1.18 [1.11,1.26]**| 1.00 [0.91,1.10]**| 1.11 [1.00,1.22]** |
| 8                                 | 1.07 [1.00,1.14]**| 1.20 [1.13,1.28]**| 0.97 [0.88,1.07]**| 1.14 [1.03,1.25]** |
| 9                                 | 1.16 [1.09,1.23]**| 1.36 [1.27,1.45]**| 1.04 [0.94,1.15]**| 1.31 [1.18,1.45]** |
| 10 (affluent)                     | 1.19 [1.12,1.27]**| 1.48 [1.39,1.58]**| 1.11 [1.01,1.23]**| 1.57 [1.42,1.74]** |
| Missing                           | 0.71 [0.59,0.85]**| 0.87 [0.73,1.05]**| 0.58 [0.44,0.76]**| 0.80 [0.59,1.07]** |

Observations
529,910
578,809
Pseudo R-squared
0.026
0.027