SUPPLEMENTARY FILE 1: DATA CLEANING

Note: all numbers are rounded to nearest 10, and all percentages to one decimal place, in compliance with statistical disclosure rules. This may result in rounding error.

The steps we followed to clean the child in need (CiN) census data as supplied by the Department for Education (DfE) are summarised in Table S1.1. More detail is provided in the text that follows.

Table S1.1. Overview of CiN data cleaning

Step	Description	Number of rows/IDs where these change*
1	Load data (and drop small number with missing IDs)	Rows: 9,752,350 Unique lachildid: 2,585,570 Unique PMR: 2,266,160
2	Check validity of start and end dates	--
3	Drop records pre-April 2008	Rows: 8,976,960 lachildid_concat: 3,742,690† Unique PMR: 2,138,920
4	Check validity of PMRs	Rows: 8,976,960 Unique lachildid_concat: 3,742,690 Unique PMR: 2,132,880
5	Check missing start dates	--
6	Check whether referral date > closure date	--
7	Check whether referral date > 31 Mar 2019	--
8	Check whether referral NFA, closure reason, ethnicity, gender and primary need code are inconsistent	--
9	Check whether NFA missing	--
10	Check end dates where referral is marked NFA	--
11	Deduplication	Rows: 8,975,420 Unique lachildid_concat: 3,742,690 Unique PMR: 2,132,890
12	Miscellaneous functions	--
13	Drop missing age and age > 17	Rows: 8,066,880 Unique lachildid_concat: 3,333,500 Unique PMR: 1,916,610 Unique id_combined: 3,170,040
14	Save final dataset	***Final dataset*** Rows: 8,066,880 Unique lachildid_concat: 3,333,500 Unique PMR: 1,916,610 Unique id_combined: 3,170,040

NFA no further action; PMR Pupil Matching Reference.

ID variables (explained in the text below): id_combined: an identifier that uses PMR where available, otherwise uses la_childid_concat; lachildid: an encrypted identifier unique within a local authority; la_childid_concat: an encrypted identifier unique within a local authority that combines the la_childid and local authority codes.

* Values for PMR ignore missing values (about 1/3 on initial load). † This increases compared to lachildid in step 1 because now we have distinguished between different LAs' use of the same ID for presumably different children. The number of unique PMRs barely changes because this is nationally unique.

SUPPLEMENTARY FILE 1: DATA CLEANING

Step 1: Load data

In this section, the raw data are loaded and some basic procedures, like converting strings into dates where relevant, are carried out.

There are two child IDs: LAchildID_anon and pupilmatchingrefanon (PMR). Almost every child has an lachildid_anon, which is an anonymised version of the ID used by LAs. Different LAs can use the same ID for different children and so at the end of this section, a concatenated lachildid_anon_concat which appends the LA code to the lachildid_anon is created. This gives unique IDs.

Not every child has a PMR. A tiny number of children have a PMR but not lachildid_anon. In these instances, the lachildid_anon is set to the PMR. There was then a tiny number of children without either PMR or lachildid_anon: these were removed from the dataset.

Step 2: Check validity of start and end dates

There are two referral dates and two closure dates labelled "cin" and "latest." We used the "cin" version as this appeared to be the actual referral/closure date per episode. Missing cinreferral/closure dates were filled in from the "latest" versions.

Step 3: Drop records pre-April 2008

Referrals starting before 1 April 2008 were removed from the dataset. Likewise, episodes ending before 1 April 2008 were dropped (or if cinclosuredate_clean was missing, the record was retained).

Step 4: Check validity of PMRs

This step draws down some PMRs into records based on lachildid_concat. This is possible where a lachildid within an LA has a PMR in a later record but not an earlier one. See Table S1.2 for a hypothetical example.

Table S1.2. Hypothetical situations where a pupil matching reference is available in later records but not earlier records for a given child.

lachildid_concat	pupilmatchingrefanon
abcde-100	NA
abcde-100	NA
abcde-100	123456ABC
abcde-100	123456ABC
abcde-100	123456ABC
abcde-555	NA
abcde-555	NA
abcde-555	NA

Records are sorted in ascending date order. Child abcde-100 is the child abcde within LA 100. Later records have a PMR but not earlier. We can "draw down" the later PMRs into the missing cells so that PMR is available for all records. Child abcde-555 does not get this PMR; although the lachildid is the same – abcde – the LA is different, so we assume this is a different child). The results of this procedure are displayed in Table S1.3.

Table S1.3. Hypothetical results from the procedure used to "draw down" a known pupil matching reference number into a child's earlier records.

lachildid_concat	pupilmatchingrefanon
abcde-100	123456ABC
abcde-100	123456ABC
abcde-100	123456ABC
abcde-100	123456ABC
abcde-100	123456ABC
abcde-555	NA
abcde-555	NA
abcde-555	NA

Step 5: Check missing start dates
Some records had a missing start date. Data are returned by local authorities to DfE on the basis of academic years. We therefore know that the record was open during the given academic year (which is never missing) and so we notionally set the referral date to the first day of the academic year.

Step 6: Check whether referral date > closure date
Where this occurs, the referral date is set to the closure date. This is because the closure date is within the academic year and so we keep the episode within the year.

Step 7: Check whether referral date > 31 March 2019
31 March 2019 is the last date possible in our extract. A tiny number of referrals had referrals greater than this date. These were set to 31 March 2019.

Step 8: Check whether referral NFA (no further action), closure reason, ethnicity, gender and primary need code are inconsistent across records for the same referral
A single referral can have several rows. This would occur where the referral is open in more than one academic year and/or where child protection plan data is added. Sometimes the referralnfa, closure reason, gender, ethnicity and primary need code variables are inconsistent between records for the same referral. The modal value (or the latest value where multimodal) across each referral's records is taken.

Step 9: check whether NFA missing
Where referralnfa is missing, its value is assumed to be 0 (i.e., there was NOT no further action on that referral or, in other words, there was in fact further action such as an assessment).

Step 10: Check end dates where referral is marked NFA
If a referral results in NFA, it is assumed that the end date is in fact the date of referral.

Step 11: Deduplication
The dataset was deduplicated on LAchildID_anon_concat, academic year, PMR, gender, ethnicity, LA, referraldate, closuredate, date of initial CPC, referralNFA,

reasonforclosure, number of previous CPPs, referral source, category of abuse, initial category of abuse, latest category of abuse, CPP start date and CPP end date.

Step 12: Miscellaneous functions

A notional_dob is calculated which is the first day of the month from the ym_birth variable (e.g., a child born 2005-11 gets a notional_dob of 2005-11-01).
This is necessary so approximate ages at referral can be calculated. After doing so, there were a number of children with negative ages. It was assumed that if the negative age is within 7 months (in fact, 7*31 days) of referral, then the referral was a pre-birth referral. Otherwise, the DOB was assumed to be erroneous. In the latter case, the notional_dob was set to the referral date and age was recalculated.

There were some children with expecteddob recorded (i.e., known pre-birth referrals). Some of these children had missing ym_birth (and hence missing notional_dob). In such cases, the expecteddob was used to infer the notional_dob (i.e., notional_dob set to expecteddob). Age was then recalculated and, again, there were ages <-7*31 days before referral. These were dealt with in the same way.

Finally, approximate age in years at referral was calculated (age_at_ref_days / 365.25).

We also carried out other routine cleaning operations such as rationalising ethnicity into five broad groups. Counters were also derived to identify the first row per child and the first row per referral per child.

We also created an ID which combined the LA child ID and PMR, using PMR where available and the LA child ID, where not. A flag was also created to indicate which ID was used in this variable.

Step 13: Drop missing age and age > 17

Records for children with a missing age or an age greater than 17 years of age at referral (i.e., 18 years or over) were removed from the dataset.