CSC data cleaning

Matthew Jay, matthew.jay@ucl.ac.uk, 18 February 2025

This document details cleaning of the ECHILD CSC data. Numbers in this documented are rounded to the nearest 10 in line with the SDC rules applicable to CSC data. An unrounded version of this document is available on the SRS with the processed data.

This work was undertaken in the Office for National Statistics Secure Research Service using data from ONS and other owners and does not imply the endorsement of the ONS or other data owners. This output has been cleared by ONS for publication (STATS20026, 08/07/24).

For CiN, assessment factors data have not yet been cleaned.

For CLA, there are two files, `CLA_2006_to_2021` and `CLA_Data_Supplied`. The file `CLA_Data_Supplied` contained 2,482,440 rows whereas `CLA_2006_to_2021` contained 1,471,670. `Data_Supplied` also contained two more variables (reason for placement change and `sample_flag`). The latter appears to be a flag for the 1/3 sample. As `Data_Supplied` therefore appeared more complete, extending to before 2006, I used that and disregarded `CLA_2006_To_2021`.

Note also that ECHILD holds a complete copy of the CLA and CiN data (i.e., not just the latest episode in the year). The CLA data also has the reason episode ceased variable.

Scripts

00_run.r – sets the working directory, defines function `mode_fun()`, loads relevant packages and executes the subsequent scripts

01_cin.r – cleans and saves the CiN data

02_cla.r – cleans and saves the CLA data

CiN data

Step	Description	N
0	Loads all CiN data from SQL server	Rows: 11,682,430
1	Cleans two of the LA variables (`CIN_LA` and `CIN_CIN_LA`), sets all var names to lowercase and removes the `cin_` prefix.	No change
2	IDs: first fixes the `lachildid_anon` as this is loaded with class ODBC_binary. Removes <10 rows where `lachildid_anon` is missing. Concatenates LA with `lachildid_anon` into new var `lachildid_anon_concat`, to ensure uniqueness. Then fills in empty PMRs where possible from a child's (based on `lachildid_anon_concat`) later records.	Rows: 11,682,430 Unique children: lachildid_anon_concat: 4,550,380 PMR: 2,576,610
3	Start and end dates: sets old start and end dates (< 1930-01-01) to missing; also sets to missing start and end dates where > 2021-03-31 as this is currently the last possible date in the dataset. Sets CPP start and end dates to NA where < 1930-01-01.	No change
4	Referral date > closure date: Where referral date is later than closure date, the closure date is set to referral date.	No change

5	Referral NFA consistent: where the referral no further action (`referralnfa`) variable is inconsistent among all records for a single referral, the modal value is taken, or where multimodal, the last value.	No change
6	NFA missing: where `referralnfa` is missing, it is set to zero (i.e., it is assumed missing means that there was not no further action).	No change
7	Closure reason inconsistent: cleans closure reason values and takes modal/latest value if inconsistent across a single referral.	No change
8	Ethnicity: the major and minor variables are available and were cleaned (e.g., setting the missing and invalid values to NA). The modal value was then taken per child of the minor group. Finally, the major groupings were manually derived from the minor.	No change
9	Gender: cleaning and modal value (per child).	No change
10	Primary need code: cleaning and modal value (per referral).	No change
11	Calculate age: Sets `notional_dob` using month and year of birth (set to 1st of month and cleaned by taking modal value per child). Approx `age_at_ref_days` and years calculated. Referrals where age at ref <0 are flagged as pre-birth referrals if they occurred within 7*31 days of the `notional_dob`. Any pre-birth referrals older than this are flagged (`dob_flag`) as potentially problematic. Note that DfE have supplied an age at referral variable. This has been renamed age_at_ref_yrs_dfe but has not otherwise been cleaned (it contains some invalid values). Year of birth variables derived.	No change
12	Derive year variables: creates year variables for referral date. Re-orders the columns. Drops the variable `seensocialworker` as this was mostly NA.	No change
13	Tidy: removes the following variables: • yearofbirth (now contained in notional_dob) • monthofbirth (now contained in notional_dob) • gender (now female) • cinat31march (can be derived from cleaned data if needed) • seensocialworker (mostly NA) • started, ended, anypoint (indicate whether a CiN period started or ended during the year or was open at any point – can all be derived from cleaned data) • servicetype, serviceprovision, serviceprovisionstartdate, serviceprovisionenddate (all are 2008/9 only)	No change
14	Referral source and disability: cleaning and modal value (per referral).	No change
15	Deduplication on all columns	Rows: 11,633,180 Unique children: lachildid_anon_concat: 4,550,380 PMR: 2,576,610
16	Assign indices: creates `row_per_child` and `epi_index` (nth episode per child)	No change
17	Identify episode types: whether an assessment was carried (`referralnfa != 1`) and whether the child was recognised as a child in need (`referralnfa != 1` and `reasonforclosure` is either	No change

	anything other than RC8 or is missing). CPP episodes are identified by virtue of having non-missing data on any of CPP start date, end date, category of abuse, initial category of abuse or latest category of abuse.	
18	Save final dataset as `cin.csv`.	No change

CLA data

Step	Description	N
1	Load data: loads all CLA data from the SQL server table CLA_2004to2019_Final. Fixes the LA child IDs as these are loaded as class ODBC_binary. Concatenates LA with `cla_child_id_anon` into new var, `la_child_id_anon_concat`, to ensure uniqueness. Then fills in empty PMRs where possible from a child's (based on `lachildid_anon_concat`) later records.	Rows: 2,482,440 Unique children: la_childid_anon_concat: 583,510 PMR: 300,820
2	Episode dates: checks no missing start and end dates. Drops <10 rows where end date < start date. No rows where start date > end date. Where `reason_episode_ceased == -10`, sets `date_episode_ceased` to NA. Derives episode calendar year.	Rows: 2,482,440 Unique children: la_childid_anon_concat: 583,510 PMR: 300,820
3	Epi start and POC start: ensures that first episode start date is the same as POC (period of care*) start date. There were 32,690 POC starts earlier than episode starts, all of which were from 2003 or earlier: these were set to the episode start date.	No change
4	Demographics: cleaning and taking modal/latest value of ethnicity, sex, year of birth, month of birth. Cleans some episode data.	No change
5	Derive year variables	No change
6	UASC status. This was coded as 1, 0 or missing. Missings were set to 0.	No change
7	Tidy and deduplicate on all variable.	Rows: 2,456,860 Unique children: la_childid_anon_concat: 583,510 PMR: 300,820
8	Calculate POC durations	No change
9	Calculate age at episode start based on derived `notional_dob`.	No change
10	Assign indices: creates `row_per_child` and `epi_index` (nth episode per child)	No change
11	Save data as `cla.csv`.	No change

* A POC is analogous to an admissions in HES: a single POC in CLA data can consist of one or more episodes. A new episode is a period of being looked after for at least 24 hours. A new episode is started if there is a change in placement status or legal status. Note that a change in the placement status may not mean a change in the actual placement/carer of a child (e.g., where a foster placement becomes a long-term placement). In order to determine whether a new episode is a change for the child, users must use the variables indicating reason episode ceased (not included in standard CLA data releases) with reason for a new episode. Users must also account for new episodes that are changes of legal status only.