Looked after children in Scotland - longitudinal data user guide

Authors: Cecilia MacIntyre, Craig Kellock, Thomas Alexis, Ross Waddell

Contents

1 Introduction

This document summarises the process of creating a longitudinal dataset based on annual returns of looked after children data supplied by local authorities (LAs) to Scottish Government (SG). The data come from LA Management Information Systems (MIS), and are collected for administrative purposes. The Administrative Data Research (ADR) Scotland programme has funded this work to create a linkable dataset which is available for research purposes, in the public benefit.

The data collected on looked after children is used by the Children and Families analysis team in Scottish Government to produce official statistics[1]. Extracts of the data have been provided to researchers subject to information governance procedures. The current ADR Scotland funded work seeks to make the data in longitudinal format available to researchers in a way that minimises the administrative burden of data extract creation on data controllers. For more information, visit the ADR UK website.

The longitudinal datasets bring together data from a number of years which will enable analysis of a child's care experience. The longitudinal data is accompanied by metadata and information on the quality of the data, which will help guide users in appropriate use of the data. The process of creating the longitudinal datasets has resulted in improvements being made to the data, by identifying issues and either resolving where possible or identifying using a data quality flag.

This document provides an overview of the outputs of the project, and outlines the key stages in the process of creating a longitudinal dataset for research purposes.

[1] Children's social work statistics - gov.scot (www.gov.scot)

2 Description of Processing and Outputs

The data in the final output combines annual returns from 2008-09 to 2018-19. The information includes details of the child in care, the episode of care which is a continuous period during which the child is legally cared for, and individual care placements or settings. Associated with each episode of care are the legal reasons in place for each care placement.

The annual returns include all children in the care of the local authority during the reporting period (1st August to 31st July)[2] and information is provided on the current episode of care, which could have started in a previous reporting period.

Two longitudinal datasets are available; one is at placement level, meaning there is one record per placement, with information from the child and episode file types attached to the records. The other dataset has the legal reason information for each child, with one legal reason start and end date per record.

2.1 Data processing

The placements and legal reasons are provided separately because the dates of placements and legal reason dates do not align in all cases, which is partly due to the information coming from different sources. A series of legal reasons is assigned to each episode of care, just like a series of placements is assigned to each episode; there should always be at least one legal reason while a child is in care but there is no requirement in the data collection for the legal reason and placement dates to be exactly aligned.

The main processing which takes place to produce the longitudinal dataset is to combine records from individual files for each year, remove duplicate records which have been supplied in more than one return and flag where information has changed between returns. In addition, some edits were applied to the data to retain the more recent information, in the case where inconsistencies were identified e.g. changes to dates of birth.

The creation of these longitudinal datasets builds on previous work, which combined annual returns from 2008-09 to 2016-17. Improvements were made following feedback from data providers and users, and as the result of additional data quality checking.

The main improvements which have been made are:

- Edits applied to original input files where there was evidence of changes in variables which would affect linkage e.g. systematic changes to social work identifier
- Improved linkage of children across dataset where a child potentially moved authority, or local authority changed identifiers

[2] Data collection for the period 2008-09 was from 1st April 2008 to 31st March 2009, and for 2009-10 was from 1st April 2009 to 31st July 2010.

- Additional detailed data quality flags to identify specific issues

2.2 Outputs

The details of the datasets can be found in the metadata published on the Research Data Scotland website. Data access instructions can be found in the metadata.

The dataset includes details on approximately:

60,000	children
70,000	episodes of care
147,000	care placements
195,000	legal reasons

These records include some outdated records which can be removed for analysis purposes. See section 4 below for information on identifying potentially outdated records.

3 Producing the longitudinal dataset

3.1 Input data - Looked after children annual returns

Data is received by Scottish Government annually from local authorities and includes information on any child in care during the reporting period. Data comes through the Scottish Exchange of Data, and is subject to initial validation checks before the data can be supplied. The Children and Families analysis team carry out additional validations before the dataset is finalised for use in the Children's Social Work Statistics publication. It is received in four file types each year: child files, episode files, placement files and legal reason files.

The child files contain identifying information about individuals such as social work identifier on LA management information systems and Scottish Candidate Numbers (SCN), provided for children who attend or have attended Scottish schools. A child level unique reference number is created by combining local authority and social work identifier, referred to as unique ID.

The episode file contains episode start and end dates, destination accommodation when a child leaves care, information on pathway plans, care plans and permanence variables (from 2016-17 onwards).

The placement files contain dates of placements in care and the placement type.

The legal reason file type contains each legal reason which was applicable during the care episode, along with the dates.

The specification and guidance documents used to provide the data to Scottish Government, and the initial validations are documented at the link below for the most recent return, and historic specifications and guidance are available on request to children.statistics@gov.scot. Scottish Exchange of Data: looked after children - gov.scot (www.gov.scot)

The main changes which occurred in the data specification and validations over the period are summarised below. The details of the changes are included in the specification for the year.

Table 1. Changes to data specification and collection guidance over time

2012-13	Legal reasons - changes to wording to match with legislations
2012-13	Collect additional support needs rather than disability
2014-15	Legal reasons - changes to wording to remove old terminology.
2015-16	Collect disability with reduced options rather than additional support needs
2016-17	Variables added • foster placement type • date on which permanence away from home was recommended • date of decision by agency decision maker • date application for a permanence order was submitted to court
2017-18	Changes to terminology on pathway plans to match with legislative changes
2017-18	Additional options for destination accommodation

3.2 Investigating data quality issues

The process of combining data from a number of annual returns provides opportunities for additional checking over and above those used at the point the data was delivered. The table below presents a summary of the issues detected in the initial exploration of the files provided by local authorities, and the action taken as part of the processing. This includes the creation of a new identifier which links children who have multiple social work identifiers in the original dataset.

Table 2. Summary of data issues in the input files

Description of issue	Number of cases affected and action
Multiple and invalid SCNs	Invalid SCNs were replaced with valid ones in later years, if one existed. Approximately 1.2% of individuals have more than one valid SCN associated with their care history.

	Approximately 0.1% of individuals had only an invalid SCN which was not later replaced by a valid one. *Action taken*: Created a variable to show the most recent valid SCN for each child
Missing SCN for linkage purposes	After the most recent valid SCN was found for each individual and matched to all records, there were 30.3% of children missing an SCN. However it is only expected that a child would have an SCN if they were aged 5 or over. There are 76.1% of children who are over 5 by the latest point they were in care in the longitudinal dataset (excluding those with missing episode end date in an earlier extract as this may be an indicator that the true end date of care is not recorded); of these 14.4% of children have no valid SCN. *Action taken*: Created a data quality flag to indicate missing SCN
Records with the same social work ID but different DOB	4% of children had more than one date of birth (DOB) recorded. If a child's date of birth changed more than once, they will be double counted in this figure. The majority of differences related to changes in recording practice by the local authority. *Action taken*: Process dates of birth to retain: i) most recent date of birth for all records except those noted in point ii). ii) For 0.6% of cases where the older DOB is more likely to be the true DOB, retain this. For these cases the most recent DOB was always the first of the month and was sent to us in a year when practice was to send partial DOB.
Records with the same SCN but different social work ID	1.6% of records with common SCNs but different social work IDs. This may relate to children moving to a different LA, but may also be related to changes in identifiers within LAs or other data entry issues. *Action taken*: Where the same gender, DOB and SCN assume records represent the same child and assign them a common identifier – referred to as correct unique ID. Additionally, some records were linked based on common SCN, gender, month and year of birth (not full DOB).

3.3 Process of creating longitudinal file

The longitudinal datasets are created by combining together the data for each year for all available years for the four file types. See flow chart 1 for an overview.

Before combining data, a new corrected unique identifier (referred to as correct unique ID) for a child was created which took account of the situation where a child had more than one social work identifier. This was caused by:

- Systematic changes made to social work identifiers by the local authority between returns
- Children being in the care of more than one local authority

In some cases the local authority was able to inform how identifiers between years were related. In other cases records were combined where they had the same SCN, gender and date of birth. It should be noted that researchers accessing the data in the National Safe Haven will only see a project specific identifier, based on the correct unique ID, this is part of the process to de-identify the data.

When annual returns are combined, there are initially many duplicate records for the same individual, because a return includes all placements for open episodes of care. For the child file, the process retains the record from the most recent return. Placements are identified as unique by having a unique combination of correct unique ID and placement start date. For the episode and legal reasons file types, the most recent occurrence of each start date is retained.

If placements which only last one day are received in a return, they are retained in the longitudinal dataset. This may be in addition to other placements which start on the same day (see section 4). This was done to ensure that legitimate placements which lasted a single day were not removed.

One of the biggest issues with the longitudinal dataset is that episodes of care are not closed in the return, but information on the child with an open placement does not continue in the subsequent year. This is mainly due to changes being made to social work systems after the return has been provided to Scottish Government. This issue has affected almost 6% of records; that is, children whose entire care histories display this issue at some point comprise approximately 6% of total records. In some cases an episode end date is missing, but the final placement end date is provided in which case it is used as the episode end date. For these records the destination accommodation is missing in approximately 72% of cases, though it is likely that the child has left care.

The dataset has also been checked for any invalid codes in variables. Table 3 below summarises any recoding that has been done as a result of these checks. It is necessary to refer to the variable codes in the metadata to understand why these steps were taken.

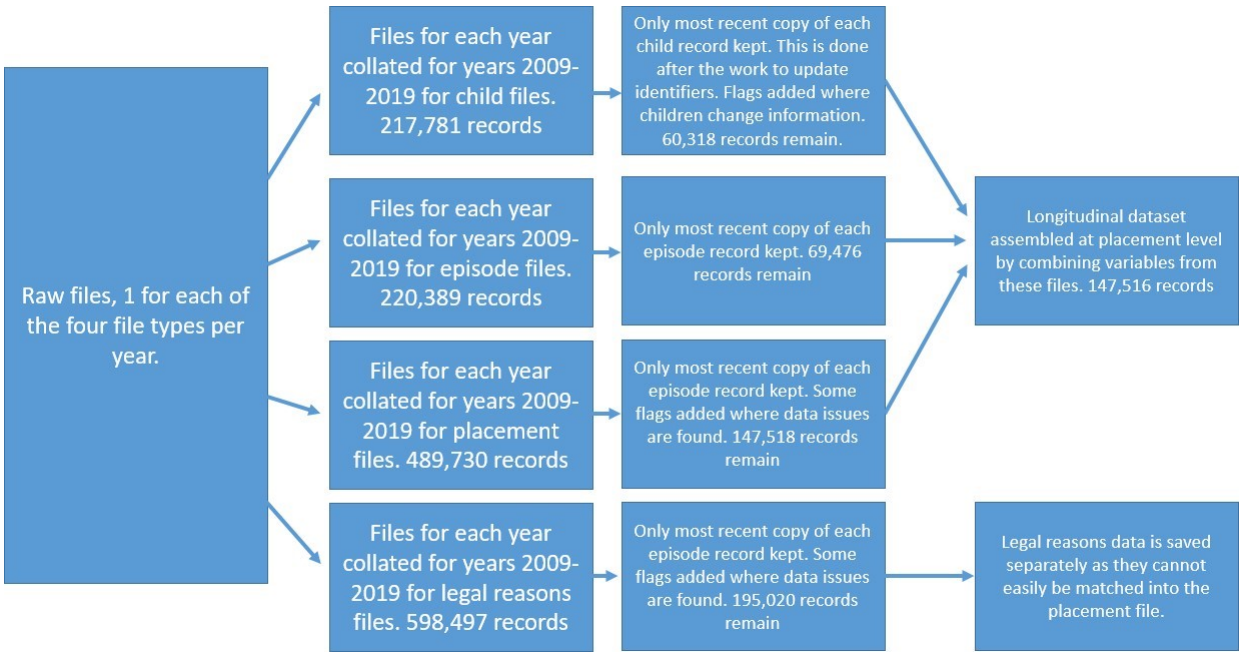
Table 3. Summary of recoding of invalid variable codes

Variable name	Recoding done
destinationaccommodation	Added a leading zero to any codes that were missing it.
ethnicgroup	Added a leading one to any records that were missing it, recoded leading zeros to ones, and changed code 05 to be 'not known' as this code was not recognised.
pathwaycoordinator	Changed code 's' to be missing.
pathwayplan	Changed code 's' to be missing.
placementtype	Added a leading zero to any codes that were missing it.
legalreason	Added a leading zero to any codes that were missing it.

As a result of the processing described here, the final dataset includes complete sets of records for episodes of care which have closed, and currently open episodes of care. However, due to retaining unique start dates, the final dataset also includes some records from previous returns where start dates have been changed. In many instances these records are likely to be outdated, but it is not possible to know this without confirmation from the local authority.

In order to identify where the data may include outdated records, a series of checks were made on the final data and a flag indicates the status of this check. Section 4 provides further details.

Flow chart 1. Creation of the longitudinal dataset



4 Overview of data quality issues

4.1 Process of creating data quality flags

This section provides an overview of the data quality flags which were produced from the final looked after children longitudinal dataset. Checks were identified to reflect requirements of the data, and a flag was created to indicate when the requirement is not met in the final dataset. Guidance is provided on how the data quality flags can be used in data analysis.

The approach of providing additional data quality flags was taken with this updated dataset, as it includes records provided by local authorities which are potentially outdated. The data quality flags highlight where there appear to be inconsistencies in the data, and allows the user to investigate which of the records to use in analysis.

A letter is associated with each data quality flag and they are combined into a number of variables depending on the nature of the issue. Most flags are at the child level, i.e. if an issue (such as an overlap between placement dates) occurs at some point in a child's care history, the whole of that child's care history will be flagged. A description of each issue, the letter denoting it, and the variable it appears in are given below.

The scale of each issue relative to the whole dataset is also given. This is so users can easily see how many records will be excluded, should they choose not to work with records with that issue. Some flags are there to inform users of issues that are not readily apparent in the longitudinal dataset as it is currently available. For example, only the most recent gender received in the returns from local authorities is retained for each individual, but if a different value for gender was received at an earlier time then this child's history will be flagged using flag Q. All flags are at child level (they flag the whole care history for a child if the data quality issue is present) except flag M and flag N, which are at placement or legal reason record level so they can be easily excluded if the data user deems necessary. These flags are at record level because it is clearer in these cases that without these particular records the data issues would not occur; for other flags, it is harder to know which records to exclude.

4.2 Data quality flags

The tables below describe the issue flagged by each of the data quality flags. The extract variable included in the dataset refers to the return from which the data is provided e.g. for the 2018-19 return, the value of the extract variable is 19 i.e. the last two digits of the year at the end of the reporting period. There are some flags which show issues that do not currently exist (no children are affected), they are included in case records in future updates have this issue.

Table 4. Placement data quality flag

Placement data quality flag description	Percentage of children affected
A Start of first episode is before date of birth – child level	0.03%
B Start of episode(s) after the first one is before the end of previous episode – child level	1.18%
C Start date of placement is missing – child level	0.00%
D End of episode is missing for an extract which is not the most recent – child level	4.75%
E Start date of first placement in an episode does not equal start date of that episode – child level	0.04%
F End date of a placement is before start date of the placement – child level	0.00%
G Within episodes, end of placement does not equal start of next – child level	0.16%
H The end date of an episode does not agree with the end date of the last placement – child level	0.10%
I The gap between the end date of an episode and the start date of next is either 0 or 1 day – child level	1.09%
J A closed episode does not have a destination – child level	3.95%
K The same episode start and end date occur in more than one extract. This means that some of the placement dates	4.17%

within that episode have changed following updated information from local authorities – child level	
L The correct unique ID is a combination of more than one unique ID. This happens when records with the same SCN, DOB and gender have different unique IDs – child level	1.26%
M Duplicate single day placements occur – record level	0.02%
N Identifies placement records which are in a duplicate episode (See Flag K). These records are likely to have been changed in a subsequent extract – record level Note this is percentage of placement records affected, as flag is at placement level	2.43%
P SCN has changed for the same unique ID – child level	5.92%
Q Gender has changed for the same unique ID – child level	0.26%
R DOB has changed for the same unique ID – child level	3.29%
S Childid has changed for the same SCN – child level	2.37%
T An earlier episode start date was received at a later point in time from the LA - child level	0.87%
U SCN is missing – child level	14.4%

Table 5. Legal reason data quality flag

Legal reason data quality flag description	Percentage of children affected
B Start of episode(s) after the first one is before the end of previous episode – child level	1.07%
C Start date of legal reason is missing – child level	0.07%
D End of episode is missing for an extract which is not the most recent – child level	8.36%
E Start date of first legal reason in an episode does not equal start date of that episode – child level	7.00%
F End date of a legal reason is before start date of the legal reason – child level	0.01%
H The end date of an episode does not agree with the end date of the last legal reason – child level	2.55%
I The gap between the end date of each episode and the start date of next is either 0 or 1 day – child level	0.67%
K The same episode start and end date occur in more than one extract year. This means that some of the legal reason dates within that episode have changed following updated information from local authorities – child level	6.58%
M Duplicate single day legal reasons occur- record level	
N Identifies legal reason records which are in a duplicate episode (See Flag K). These records are likely to have been changed in a subsequent extract – record level	

Note this is percentage of legal reason records affected, as flag is at legal reason level.	
T An earlier episode start date was received at a later point in time from the LA – child level	

4.3 Suggestions of use of data quality flags

Table 6 gives an indication of the impact of some of the data quality flags on the dataset. This may be useful carrying out preliminary analysis of the dataset as it contains outdated records. Tables 7 and 8 give suggestions for how each flag can be used.

Table 6. Working with data quality flags – an example

Steps	Impact
Remove all placements which are included in duplicate episodes – flag N	This reduces the file from 147516 to 143932. 59507 children
Create a dataset which contains children where they have no data quality flags for (B, E, F, G and H) which are the main ones which are flagged by out of date records	This reduces the file to 139982 records 58687 children
Resolve the remaining issues if relevant	For example 2532 cases with a missing end date, and 600 with no gap between episodes.
Carry out preliminary analysis of cleaned data and investigate an approach to the outstanding data issues	

Table 7. Placement data quality flag – suggestions for use

Placement data quality flag description	
A Start of first episode is before date of birth	Drop child as likely to be error in event dates or date of birth. A possible explanation for this is children who are looked after who are also pregnant. There is some confusion about whose DOB to record, and this could explain why a child is recorded as looked after before they are born. Also the DOB may be recorded initially as the expected delivery date for pregnant mothers. When the actual DOB is recorded this will appear to change.
B Start of episode(s) after the first one is before the end of previous episode	Examine event dates for possible out of date records which have been changed. These records could be removed.
C Start date of placement is missing	Drop child as likely to be error in event dates
D End of episode is missing for an extract which is not the most recent	Make assumption that end of episode is in previous extract. The end dates for these episodes could be defaulted to the end of the extract period in which they occur as an estimate of when the child left care.
E Start date of first placement in an episode does not equal start date of that episode	Examine event dates for possible out of date records which have been changed. These records could be removed.
F End date of a placement is before start date of the placement	Examine event dates for possible out of date records which have been changed. These records could be removed.
G Within episodes, end of placement does not equal start of next	Examine event dates for possible out of date records which have been changed. These records could be removed. Placement dates are linked to LA payment systems, there may be overlaps where they have paid foster carers for a full day even where they only looked after the child for half a day.

H The end date of an episode does not agree with the end date of the last placement	Examine event dates for possible out of date records which have been changed. These records could be removed.
I The gap between the end date of an episode and the start date of next is either 0 or 1 day	Combine sets of placements into one episode. Feedback from LAs has been that this may occur as a result of the separation of systems for recording placements and legal reasons. It is likely a recording error when this occurs.
J A closed episode does not have a destination	Imputation of missing destination
K The same episode start and end date occur in more than one extract. This means that some of the placement dates within that episode have changed following updated information from local authorities.	For these children, the placements in the duplicated episodes should be removed. These are flagged in Flag N
L The correct unique ID is a combination of more than one unique ID. This happens when records with the same SCN, DOB and gender have different unique IDs.	Flagged for data quality information. Could indicate the child has moved to a different LA, or that the LA has changed their identifiers.
M Duplicate single day placements occur	Flagged for data quality information. This could be related to the data that comes from LA finance systems; when two carers are paid for a full day each when they each looked after a child for half the day each for example.
N Identifies placement records which are in a duplicate episode (See Flag K). These records are likely to have been changed in a subsequent extract. Note this is percentage of placement records affected, as flag is at placement level.	Filter out these records.
P SCN has changed for the same unique ID	Flagged for data quality information. Different identifying information has been received from the LA for this unique ID in the past.
Q Gender has changed for the same unique ID	Flagged for data quality information. Different identifying information has been received from the LA for this unique ID in the past.

R DOB has changed for the same unique ID	Flagged for data quality information. Different identifying information has been received from the LA for this unique ID in the past. This may be related to unaccompanied asylum seekers. It is not until the young person is actually in the LA that their full DOB is recorded, they just get an initial approximate date at first.
S Childid has changed for the same SCN	Flagged for data quality information. Different identifying information has been received from the LA for this SCN in the past.
T An earlier episode start date was received at a later point in time from the LA	Examine records for possibility that outdated records exist in this child's care history and remove.
U SCN is missing	Flagged for data quality information. All children old enough to go to school should have an SCN.

Table 8. Legal reason data quality flag – suggestions for use

Legal reason data quality flag description	
B Start of episode(s) after the first one is before the end of previous episode	Examine event dates for possible out of date records which have been changed. These records could be removed.
C Start date of legal reason is missing	Drop child as likely to be error in event dates
D End of episode is missing for an extract which is not the most recent	Make assumption that end of episode is in previous extract, and is the same as in placement file. The end dates for these episodes could be defaulted to the end of the extract period in which they occur as an estimate of when the child left care.
E Start date of first legal reason in an episode does not equal start date of that episode	Examine event dates for possible out of date records which have been changed. These records could be removed.
F End date of a legal reason is before start date of the legal reason	Examine event dates for possible out of date records which have been changed. These records could be removed.
H The end date of an episode does not agree with the end date of the last legal reason	Examine event dates for possible out of date records which have been changed. These records could be removed.
I The gap between the end date of each episode and the start date of next is either 0 or 1 day	Combine sets of legal reasons into one episode, using same approach as placement records.
K The same episode start and end date occur in more than one extract. This means that some of the legal reason dates within that episode have changed following updated information from local authorities.	For these children, the placements in the duplicated episodes should be moved. These are flagged in Flag N
M Duplicate single day legal reasons occur	Flagged for data quality information
N	Filter out these records.

Identifies legal reason records which are in a duplicate episode (See Flag K). These records are likely to have been changed in a subsequent extract. Note this is percentage of placement records affected, as flag is at placement level.	
T An earlier episode start date was received at a later point in time from the LA	Examine records for possibility that outdated records exist in this child's care history and remove.

Document details

Version	Date published	Changes
1.0	17/06/2022	N/A

Produced by ADR Scotland

ADR Scotland brings together specialist researchers and statisticians from Scottish Government's Data for Research Unit and Scottish Centre for Administrative Data Research (SCADR). ADR Scotland is part of Administrative Data Research UK, funded by the Economic and Social Research Council.