**Deliveries cohort**

8 August 2025, Matthew Jay, [matthew.jay@ucl.ac.uk](mailto:matthew.jay@ucl.ac.uk)

**About this document**

This read me contains information about the files that comprise the deliveries cohort, the cohort itself, its data dictionary, the R scripts used to create the cohort and some guidance on working with the cohort.

**Files**

The HES deliveries cohort consists of a core deliveries file (deliveries\_cohort\_full.csv), two files containing the cohort members’ diagnostic and procedure coding data across HES APC (diagnoses\_long.csv and operations\_long.csv), a data dictionary (data\_dictionary.xlsx) and a series of code lists, most of which were taken from the ECHILD Phenotype Code List Repository (<https://code.echild.ac.uk>). The current version of the dataset was compiled between 04/08/25 and 07/08/25.

**Core deliveries file**

The HES deliveries cohort contains 15,129,641 deliveries identified in HES for 8,759,868 mothers. Of the 15 million deliveries, 7,200,860 represent first live births (to 7,200,860 mothers). Included are all deliveries recorded in HES from 1 April 1997 to 31 March 2023, with the following exclusions were applied:

1. Sex was not coded 2 (female)\*
2. Mother aged <15 or >50 years at start of delivery episode
3. Mother was not living in England at delivery
4. Mother had a date of death recorded prior to her delivery date
5. Mother had HES APC activity following a recorded date of death
6. The difference between the episode start date and calculated delivery date was greater than 100 days
7. Mother had any ostensible delivery episodes that could not be assigned to a delivery.
8. The delivery had an abortion code recorded.
9. It was not possible to infer a missing numbaby (number of babies on delivery) value

\* Because only episodes where sex was coded as 2 were initially extracted, a sex variable would be a constant and is therefore not included in the dataset.

Deliveries were identified with reference to a range of variables including diagnoses and operation codes, delivery method and non-missing maternity tail variables. In most cases, the initial data extraction returned sole episodes for deliveries. However, there were some women with multiple ostensible delivery episodes close to each other.

This issue was resolved by specifying a delivery window from the date of the first episode. All episodes falling within 150 days of the first episode were within that window. The episode in this window that was most likely the delivery episode was selected based on an algorithm described below. Once the most likely delivery episode was identified, the delivery date was calculated using epistart, unless there was other information such an opdate associated with a relevant delivery procedure code. Finally, a new window was set as the delivery date plus 168 days: any subsequent delivery episodes within this window were ignored. Subsequent deliveries must have been beyond this window. There were some women with potential delivery episodes between the first episode start date + 150 days and the delivery date + 168 days. These were the women with possible delivery episodes that could not be assigned to a delivery and were excluded.

**Data dictionary**

The data dictionary explains each variable. Where possible, coding as used in the HES APC TOS is retained. Otherwise, variable coding should be obvious (e.g., dates or binary variables).

**Diagnostic and procedure coding**

These two files contain, in long format, non-missing ICD-10 and OPCS-4 codes from all APC records for all cohort members. The core cohort file does contain some phenotyping data (certain phenotypes recorded in the 3 years prior to birth), however, this may not be sufficient for all users’ needs. The diagnoses\_long.csv and opertions\_long.csv files were created to facilitated custom phenotyping.

**Old delivery cohort**

This delivery cohort is based on a previous version that was developed primarily in SQL by Georgina Ireland. Because the code for that cohort was not automated or easily reproducible, it was recreated, making some enhancements (including almost full automation) along the way. The exact contents of this cohort may therefore differ from previous versions.

**R scripts**

Using the R scripts, updating the cohort in future should require minimal changes to the code (e.g., updating the years of data to be extracted) and then monitoring to ensure refreshed data do not cause errors. While the intention is that most users will not need to inspect the code that created the cohort, it is documented here to make updating and modifying it in future easier.

You will see that at the end of each script, a temporary dataset is saved in \*.rda format. These files are deleted in the final script and only the final dataset is saved in \*.csv format for faster reading and compatibility across software packages.

|  |  |
| --- | --- |
| Script | Description |
| 00a\_run\_scripts.r | Sets working directory, loads libraries, connects to the SQL server, specifies cohort inception years and runs all scripts. |
| 00b\_functions.r | Loads various functions. |
| 01\_extract\_delivery\_episodes.r | Extraction of delivery episodes for the given delivery years. |
| 02\_initial\_cleaning.r | Carries out basic cleaning functions like specifying system missing values, ensuring dates are class Date and setting impossible antedur and postdur values to missing. |
| 03\_initial\_processing.r | Creates flags for still births and multiple births (though we return to these later once deliveries and maternity tail variables are cleaned) and calculates the completeness of the maternity tail (i.e., number of non-missing maternity tail variables), which is used in identifying the correct delivery episode in the next script. |
| 04\_derive\_delivery\_episode.r | In most instances, it is clear which HES episode is the delivery episode because most women only have one HES episode that could be delivery episode. Either it is the sole delivery and sole episode, or the women has multiple deliveries initially extracted (in script 01), all of which are sufficiently far apart in time (in our case, more than 150 days) that there is no doubt that they are indeed separate deliveries.  Some women, however, have several episodes in the initial extraction which are close to each other (within 150 days of the earliest episode). This is because providers may code information on a maternity-relevant episode but which is not actually the delivery. In these instances, it is necessary to select which episode is most likely the delivery episode.  The algorithm in this script starts by identifying the earliest episode start date per woman and defines a window of this date plus 150 days. If there is only one episode within this window (i.e., the earliest), then this is deemed the delivery date. If there is more than one, the most likely episode is selected in the following order.   1. The episode with the most complete maternity tail record. This only applies where there is one episode with the highest number of maternity tail variables complete (i.e., if there are two equally complete episodes, we move to the next step). 2. The unique episode with the lowest antedur (i.e., days admitted before delivery) value. 3. The unique episode with the lowest postdur (i.e., days admitted after delivery). 4. The unique episode with the greatest number of delivery procedure codes. 5. The earliest episode that appears to be the mother’s first live birth (if there are multiple with the same earliest epistart, the min of epikey is used). 6. In all other cases, the episode with the earliest epistart.   The algorithm then specifies the delivery date, which is set using antedur, postdur and delivery procedure code dates, where available, else the epistart. Finally, a new window is set as the delivery date plus 168 days. Only episodes beyond this window are counted as potential delivery episodes when identifying subsequent deliveries.  The algorithm works in a loop through all possible deliveries. Because the maximum number of deliveries is not knowable in advance, the algorithm begins to check whether there are any remaining unassigned deliveries after the 17th loop. If it finds no new deliveries, the loop is terminated.  The identified deliveries are then combined into one file and the dataset is saved as a temporary dataset. |
| 05\_unassigned\_episodes.r | Finds the episodes that could not be assigned to a delivery because they fell between the earliest episode start date + 150 days and the delivery date + 168 days. These are used later as an exclusion criterion. |
| 06\_get\_ethnicity.r | Extracts ethnicity data across all APC records, takes the modal value and merges this into the cohort data table. |
| 07\_clean\_maternity\_tail.r | Further cleaning of the maternity tail, in particular identifying singletons, twins and multiples. Numbaby is cleaned as per the rules documented in the data dictionary. |
| 08\_birth\_status.r | Birth status (all live born, some live born, all still born) is derived. |
| 09\_numpreg\_birthorder.r | Number of previous pregnancies is cleaned and the first live birth per mother is identified. |
| 10\_get\_apc.r | Extracts diagnoses and procedures across all APC records in order to create files that contain: diagnoses codes, procedure codes, mortality identified in APC and the last APC activity. We return to the latter two in script 12. |
| 11\_implement\_phenotypes.r  11a\_ari.r  11b\_srp.r  11c\_pearson.r  11d\_hardelid.r  11e\_charlson.r  11f\_id\_sheehan.r  11g\_disability\_grant.r  11h\_enmmoi.r  11i\_high\_risk\_pregnancy.r  11j\_other.r | Implements a range of phenotypes, mostly from the ECHILD Phenotype Code List Repository:   1. Adversity-related admissions (Herbert et al) 2. Stress-related presentations (Ní Chobhthaigh et al) 3. Mental health and behavioural problems (Pearson et al) 4. Charlson comorbidity index (Quan et al) 5. Intellectual disability (Sheehan et al) 6. Disability (Grant et al) 7. The English Maternity Morbidity Outcome Indicator (Nair et al) 8. High risk pregnancy (Langham et al) 9. Others:    1. Abortion    2. Pre-term birth    3. Teenaged mother (current delivery and ever)   The code list files are all available with the cohort files for reference.  Phenotypes a to h were identified in APC records in the 3 years prior to birth. This can be altered in future updates by amending the relevant line of code in script 11. Users may wish to implement the above phenotypes over other time periods or other phenotypes. To facilitate this, the diagnosis and procedure codes across cohort members’ APC records are available. |
| 12\_get\_mortality.r | Extracts ONS mortality data and merges into main data frame date of death along with all recorded causes of death. Also merges date of death identified in HES APC records. The date of death according to ONS and APC are both available, as well as a combined date of death variable, which uses the ONS date, where it is available, otherwise the HES APC date. The ONS date is preferred because cause of death is only captured in the ONS data.  Mothers are flagged where they have any HES APC activity after their combined date of death. |
| 13\_get\_region.r | Extracts geography variables in order to identify deliveries to mothers not born in England. |
| 14\_exclusions.r | Excludes deliveries who meet any of the following conditions:   1. Aged under 15 2. Mother not resident in England 3. A date of death is indicated before delivery date 4. There is HES APC activity after a date of death 5. The difference between the episode start date and the delivery date is greater than 100 days 6. The mother has unassigned episodes 7. There is an abortion code 8. Numbaby (number of babies on delivery) is missing after cleaning   These mothers are dropped from the dataset. |
| 15\_tidy\_and\_save.r | The dataset is ordered by tokened and delivery date and then the columns are re-ordered. Variables that are actually constants (e.g., sex, abortion, excluded) are dropped. The dataset is then saved as \*.csv. The diagnoses and procedures are saved as \*.csv as is the list of column names for populating the data dictionary. Temporary files are all deleted from the hard disk. |

**Guidance on working with the cohort**

To be completed. Key points:

* Best to use data.table and fread() to load the data.
* Can derived relevant year variables from the actual dates. These do not come with the dataset to save space.
* Exclusions already dropped.
* Can identify still births and multiples with birth\_status and numbaby\_clean.
* First live birth identified with first\_live\_birth
* Death indicated by non-missing date of death variable. Cause only available where derived from ONS data
* Diagnoses, procedures and other episode data are for the delivery episode itself.
* Phenotyping can be done easily using the long diagnoses and procedure data and the code lists from the ECHILD Phenotype Code List Repo (or using those scripts as a template).