Medicines & Healthcare products Regulatory Agency

CPRD

Hospital Episode Statistics (HES)

Admitted Patient Care and

CPRD primary care data

Documentation (set 22/January 2022)

Version 2.8

Date: 20 Dec 2021

NIHR | National Institute for Health Research

Medicines & Healthcare products Regulatory Agency

CPRD

Documentation Control Sheet

Over time, it may be necessary to issue amendments or clarifications to parts of this document. This form must be updated whenever changes are made.

Version	Affected Areas Summary of Change	Prepared By	Reviewed By
1.0	Initial Draft		
1.1	Modified	Shivani Puri	Susan Eaton, Arlene Gallagher
1.2	Modified	Shivani Puri	Rachael Boggon
1.3	Modified	Alan Barcroft	Shivani Puri, Susan Eaton
1.4	Modified	Tarita Murray-Thomas	Helen Strongman
1.5	Formatted	Grant Lee	Helen Strongman
1.6	Modified	Tarita Murray-Thomas	Shivani Padmanabhan
1.7	Modified	Tarita Murray-Thomas	Helen Strongman
1.8	Modified	Helen Strongman	Shivani Padmanabhan
1.9	Modified	Tarita Murray-Thomas	Dan Dedman
2.0	Modified and Formatted	Dan Dedman, Arlene Gallagher	Tarita Murray-Thomas
2.1	Modified	Dan Dedman	Tarita Murray-Thomas
2.2	Modified	Tarita Murray-Thomas	Arlene Gallagher
2.3	Modified	Dan Dedman	Tarita Murray-Thomas
2.4	Modified	Helen Booth	Dan Dedman
2.5	Modified	Rachael Williams	Tarita Murray-Thomas
2.6	Modified	Tarita Murray-Thomas	Mike Lonergan
2.7	Modified	Susan Hodgson	
2.8	Modified	Tarita Murray-Thomas	Chisomo Mutafya

Summary of Changes

Version 1.1

- Refined wordings

Version 1.2

- Created separate data dictionary/specification for Integrated, Basic and Full HES
- Amended section on HES data and CPRD GOLD to reflect what linked data represents

Version 1.3

- Amended ‘What are the HES?’ section, including information on the HSCIC

Version 1.4

- Updated for set 10
- Updated the document title to change the focus to Admitted Patient Care data
- Updated ‘HES data and CPRD GOLD’ to ‘HES Admitted Patient Care data and CPRD GOLD’
- Removed reference to linked HES inpatient data being the only HES data source currently available in CPRD
- Added information about the match_rank variable which is newly available for set 10
- Removed reference to HES Outpatient (OP) data under ‘Future plans’ as OP data is now available as an additional data module and has its own documentation

Medicines & Healthcare products Regulatory Agency

CPRD

Version 1.5
- Formatted with new agency branding and updated document title
- Included version of HES on front page

Version 1.6
- Updated for set 11
- Clarified the information relating to the 'match_rank' variable under 'HES Admitted Patient Care data and CPRD GOLD'
- Updated section on ethnicity data derived by CPRD as recorded under 'Data structure and formatting'
- Updated the last year of collection of augmented care period data as recorded under 'Data structure and formatting'

Version 1.7
- Updated for set 12
- Removed reference to HES Outpatient and Accident & Emergency data
- Added table of proportion of patients linked by match_rank
- Added details about availability of records with match_rank values of 6 to 8
- Added details about availability of records with multiple HESIDs
- Added information under 'Known issues' relating to provisional release of HES data

Version 1.8
- Updated for set 13
- Added explanation of changed definition of the derived ethnicity variable
- Updated references to reflect change of name from HSCIC to NHS Digital

Version 1.9
- Updated for set 14
- Updated web links

Version 2.0
- Updated for set 15
- Updated header and footer with new agency branding
- HRG variable changes detailed

Version 2.1
- Updated for set 16
- Updated link address for NHS Digital HES data dictionary
- Updated to include CPRD Aurum
- Updated table numbers

Version 2.2
- Updated for set 17
- Updated document version number, date and linkage set
- Updated the HES APC coverage dates for this release
- Removed reference to integrated, basic and full HES

Version 2.3
- Updated for set 18: version number, date, linkage set, coverage dates
- Updated information on n_patid_hes and ICD-10 codes under "Data structure and formatting"
- Updated NIHR logo

Version 2.4
- Updated for set 19: version number, date, linkage set, coverage dates

Version 2.5
- Updated for set 20: version number, date, linkage set, coverage dates
- Updated to indicate that the single ethnicity variable (gen_ethnicity) in the HES_patient file has been carried from set 19.

Version 2.6

- Updated for set 21: version number, date, linkage set, coverage dates, added DOIs

Version 2.7

- Updated to revise reference to ISAC

Version 2.8

- Updated for set 22: version number, date, linkage set, coverage dates, added DOIs

Medicines & Healthcare products Regulatory Agency

CPRD

HES Admitted Patient Care (APC) data linked to CPRD primary care data

This document provides an overview of the HES Admitted Patient Care (HES APC) data, and the available subset that is linked to CPRD GOLD and CPRD Aurum.

What are the HES Admitted Patient Care data?

The Hospital Episode Statistics (HES) represent a data warehouse of English NHS data related to health care provider activity across three main patient groups:

- Admitted patient care – inpatient and day case admissions to hospital
- Outpatient appointments and attendances
- Accident and Emergency attendances

The HES data are managed by NHS Digital (https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/hospital-episode-statistics), formerly known as the Health and Social Care Information Centre. The data are extracted from a data warehouse on a monthly basis. At the end of the fiscal year there is a "month 13" annual refresh which corrects known data quality issues prior to locking the annual published data.

The HES APC data contains details of all admissions to *English* NHS health care providers. The patients include private patients and residents outside of England, who were treated by NHS health care providers, including treatment by the independent sector, if funded by the NHS. All NHS health care providers in England, including acute hospital trusts, primary care trusts and mental health trusts provide data. The data is available at the person level as a consultant episode for admitted patients.

There are extensions to the admitted patient care data that cover maternity and adult critical care (referred to as either Augmented Care Periods or as the Critical Care Minimum Data Set; as the underlying data standards have changed over time).

Data have been collected for admitted patient care data from 1989 onwards. CPRD only links data from 1997 due to the introduction of the NHS number which is an important element in the linkage of the data. More than 17 million consultant episodes are added each year. The data are recorded for episodes ending from April 1st to the following March 31st each year, corresponding to NHS fiscal years.

Before requesting HES APC data, users are encouraged to familiarise themselves with the content of HES APC data. Details on the fields available and changes to field definitions over time can be found at: https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/hospital-episode-statistics/hospital-episode-statistics-data-dictionary. Details of HES APC activity statistics can be found at: https://digital.nhs.uk/data-and-information/publications/statistical/hospital-episode-statistics-for-admitted-patient-care-outpatient-and-accident-and-emergency-data

Medicines & Healthcare products Regulatory Agency

CPRD

Accessing HES Admitted Patient Care data linked to CPRD GOLD and CPRD Aurum

HES APC data can only be accessed as part of a data extract linked to CPRD primary care data (CPRD GOLD or CPRD Aurum). Access is provided by CPRD subject to protocol approval.

Not all patients in CPRD GOLD or CPRD Aurum are eligible to be linked to HES, for example, due to the region in which they reside (outside England), or the lack of a valid NHS identifier. Source files (linkage_eligibility.txt) are provided to allow researchers to identify the subset of patients who are eligible.

Linkage coverage period

The latest release of HES APC data linked to CPRD primary care data (set 22) covers the period **April 1997 – March 2021 inclusive**. Data up to March 2020 are final. Data up to March 2021 are provisional.

Linkage algorithm and the match_rank variable

The linkage between HES APC and CPRD primary care data uses an eight-step deterministic linkage algorithm based on four identifiers, shown in Table 1 below. The linkage is undertaken by NHS Digital, acting as a trusted-third-party, on behalf of CPRD. No personal identifiers are held by CPRD, or included in the CPRD GOLD, CPRD Aurum, or linked HES APC data.

Table 1: NHS Digital 8 step linkage algorithm

Step	Match
1	Exact NHS number, sex, date of birth (DOB), postcode
2	Exact NHS number, sex, DOB
3	Exact NHS number, sex, postcode, partial DOB
4	Exact NHS number, sex, partial DOB
5	Exact NHS number, postcode
6	Exact sex, DOB and postcode (where NHS number does not contradict the match, the DOB is not 1st of January & the postcode not on the communal establishment list)
7	Exact sex, DOB and postcode (where the NHS number does not contradict the match and the DOB is not 1st of January)
8	Exact NHS number

The matching steps are applied sequentially. If a CPRD GOLD or CPRD Aurum patient record is matched in one step, it is no longer available for matching in subsequent steps. Matching results are summarised in Table 2A and 2B below.

Table 2A: Number and proportion of **CPRD GOLD** patients matched to a HES patient* at each step of the linkage algorithm in set 22.

Linkage step (match_rank)	Frequency	Percent
1	5,711,965	68.6%
2	2,319,661	27.9%
3	13,432	0.2%
4	18,103	0.2%
5	3,433	0.0%
6	234,748	2.8%
7	14,437	0.2%
8	6,518	0.1%

*includes patients in all HES datasets (Admitted patient care, Outpatient, A&E & DID)

Table 2B: Number and proportion of **CPRD Aurum** patients matched to a HES patient* at each step of the linkage algorithm in set 22.

Linkage step (match_rank)	Frequency	Percent
1	23,803,029	65.6%
2	11,106,671	30.6%
3	49,717	0.1%
4	75,969	0.2%
5	13,411	0.0%
6	1,135,874	3.1%
7	65,743	0.2%
8	33,065	0.1%

* includes patients in all HES datasets (Admitted patient care, Outpatient, A&E & DID)

CPRD provides users with a match_rank variable which corresponds to the step at which a match was established. In general, a lower value for the match_rank is considered stronger evidence for a positive match. Note that only patients with a match_rank of 5 or less are considered definitive matches and are included in the linked HES APC dataset. Patients matched on steps 6-8 have been retained in separate files. We envisage that the retained records will primarily be of interest to methodological researchers. If you are interested in these data, please speak to a member of the CPRD Observational Research team prior to submission of your protocol.

A minority of patients are linked to multiple HESIDs. These patients are removed from the HES APC dataset. However, the data have been retained by CPRD and are available on request. If you are interested in these data, please speak to a member of the CPRD Observational Research team prior to submission of your protocol.

As far as possible, the linked HES APC data is supplied “as is”, without any modification or cleaning during processing by CPRD. Where CPRD has modified the HES data, these are detailed below.

Medicines & Healthcare products Regulatory Agency

CPRD

DOI

Please cite in any publications using these data:

CPRD GOLD HES APC January 2022 (set 22) - https://doi.org/10.48329/fagm-ez75

CPRD Aurum HES APC January 2022 (set 22) - https://doi.org/10.48329/vagx-9d96

Data structure and formatting

The data has been arranged into files relating to hospitalisations (alternatively known as spells in HES), episodes, and files for events that are linked to specific episodes.

Hospitalisations refer to the total period of inpatient hospital stay from admission to discharge. When a hospitalisation spans the end of the HES year, it is artificially modelled as two hospitalisations, from admission to end of HES year (in the first year's HES data) and from start of the HES year to final discharge (in the second HES year).

An episode is a time-period within a hospitalisation, which corresponds to the period where the patient is in the continuous care of one consultant using the beds of one health care provider. Note that this is not always the same as a single stay in hospital, because a patient may be transferred from one consultant to another during their stay. In these cases, there will be two or more-episode records for the hospitalisation. Consultant episodes will also terminate when a patient is transferred between health care provider organisations, even though their inpatient care may be continuous.

Each patient may have one or more HES hospitalisations. Each hospitalisation can consist of one or more episodes. For each episode, up to 20 diagnoses and 24 procedures may be recorded. Additionally, each episode can have up to nine periods of augmented care. If the HES hospitalisation is related to pregnancy, each episode can additionally have information on up to nine babies to accommodate multiple births.

For each patient cohort, HES APC data will be provided as separate text tab delimited files. These files can be linked to the corresponding CPRD GOLD or CPRD Aurum patient cohort file using the CPRD generated encrypted patient key (patid). Files can be imported into statistical software such as Stata or SAS, or into data management packages such as Microsoft Access, for further data processing and analysis.

The format of the HES data has been modified for linked patients in the following ways:

- **Unique HES patient key (gen_hesid):** A patient key has been generated to identify a unique patient in the HES data. This is unique across all CPRD-linked HES datasets including HES admitted patient care, HES outpatient and HES accident and emergency (A&E) data. An individual that has contributed data to more than one CPRD practice will have the same patient key (gen_hesid) in the HES_patient file but this may change between linkage sets. Researchers will need to consider how this may impact their study.

- **Multiple CPRD GOLD or CPRD Aurum patient records linked to a single HES record (n_patid_hes)**: In CPRD GOLD and CPRD Aurum, individuals are assigned a new patient identifier (patid) each time they move from one contributing practice to another. A single individual can therefore be represented by two or more different patient identifiers, and each can be linked to the same HES patient record. This information is captured by the variable **n_patid_hes** which indicates

how many different patient identifiers *within the same database* (i.e. CPRD GOLD or CPRD Aurum) are linked to the same HES patient record.

In a very small number of instances (affecting <0.03% of all patients), the number of primary care patient records linked to the same HES patient record is large (n_patid_hes>20). This may occur if data from two or more individuals are incorrectly allocated the same HES patient key (HESID). Although this is a very rare occurrence, **the linked HES data for these patients may not be reliable.** A simple approach when using linked HES data would therefore be to exclude from any analyses all patients where n_patid_hes is deemed to be large.

- **ICD-10 diagnosis codes:** Clinical diagnoses are coded using the International Classification of Disease 10th revision (ICD-10). ICD-10 uses a 3-character code (format *Ann*) as the primary classification of diseases, signs and symptoms, abnormal findings, and external causes of injury. An optional 4-character code (*Ann.n*) allows for specificity regarding the cause, manifestation, location, severity and type of injury or disease.
  ICD-10 allows the extension of 4-character codes by adding further characters. The adoption of these codes and the coding conventions appears to vary between providers. For this reason, any extra characters after the first 4 characters (A*nn.n*) are stored by CPRD in a separate field (ICDx). When requesting linked HES data from CPRD based on specific ICD-10 codes, it is recommended that researchers provide 3- or 4- character codes only to identify events in HES.

- **Ethnicity:** Ethnicity (ethnos) is recorded in each episode of the original HES data and these are recoded (see Table 3 below) and provided in the HES episodes table (hes_episodes.txt). Most patients have the same ethnicity grouping for each episode. However, for a minority of patients, recording of ethnicity varies between episodes, both within and across hospitalisations. CPRD use the following stepwise process to derive a single ethnicity variable (gen_ethnicity) for each subject in the patient file:
  1. The variable is set to the most frequently recorded ethnicity value across episodes and hospitalisations in the HES Admitted Patient Care, HES Outpatient and HES A&E data.
  2. Where the most frequently recorded ethnicity is “Unknown”, “Unknown” values are removed, and the value is reset to the most commonly recorded ethnicity.
  3. Where there is no majority, the derived ethnicity is recorded as “Unknown”.

Recoded Ethnicity	*Original Ethnicities*
White	0 = White, A = British (White), B = Irish (White), C = Any other White background
Black_Caribbean	1 = Black – Caribbean, M = Caribbean (Black or Black British)
Black_African	2 = Black – African, N = African (Black or Black British)
Black_Other	3 = Black – Other, P = Any other Black background
Indian	4 = Indian, H = Indian (Asian or Asian British)
Pakistani	5 = Pakistani, J = Pakistani (Asian or Asian British)
Bangladeshi	6 = Bangladeshi, K = Bangladeshi (Asian or Asian British)

Other_Asian	L = Any other Asian background
Chinese	7 = Chinese, R = Chinese (other ethnic group)
Mixed	D = White and Black Caribbean (Mixed), E = White and Black African (Mixed), F = White and Asian (Mixed), G = Any other Mixed background
Other	8 = Any other ethnic group, S = Any other ethnic group
Unknown	9 = Not given, X = Not known, Z = Not stated

Table 3: Ethnicity recoding by CPRD

- **Hospitalisations (within a health care provider):** A hospitalisation level file was created, containing the spell number (uniquely identifying a hospitalisation), dates of admission and discharge, and duration of hospitalisation (in days).

- **Augmented Care Data**: This area has been noted by HES as having some data quality issues. We have included the data mostly "as is" except structuring it into a separate file. The limitations reflect that some hospitals record augmented care periods using systems which may not show up as augmented care in the HES data. There can be up to nine augmented care periods during a single episode. Since augmented care focuses on keeping patients alive, there can be overlapping episodes (where multiple 'teams' have a role at the same time). This means the numbers of days in augmented care do not always correspond to the number of days within an episode. The variable 'numacp' determines the number of augmented care records per episode. Augmented care data is available until the year 2007/2008, after which it has been replaced by the Critical Care Data.

- **Critical Care Data**: The source of HES critical care data is the CCMDS (Critical Care Minimum Data Set), which includes records for critical care periods in adult designated wards. Any one patient can have multiple critical care stays, which may be in the same or different time period for the same or different condition. Critical care data is available from HES years 2008 onwards.
  We are aware that patid and epikey do not always uniquely identify an episode of care in the Critical Care Data, that is, for some patid-epikey combinations there may be multiple records with different data for the same episode of care. In such situations, it is not possible to determine which record to retain. Depending on the requirements of the study, it may be necessary to exclude patients with multiple record combinations from the study or to exclude the affected records only. In some cases, excluding or retaining only those data variables needed for the study may help to resolve issues of multiple patid-epikey combinations e.g. in an analysis of counts of ICU admissions, keeping only one patid-epikey-date combination from among multiple records of admissions with the same date but different data may avoid issues of multiple patid-epikey combinations.

- **Maternity data:** This area has also been noted by the HES as having some data quality issues. As with the augmented care data, these data are restructured into a separate file in an array format. There can be information recorded on up to nine births within a single episode (six births for years prior to 2003). Otherwise, the data has not been altered. Several quality issues may be readily obvious. Two variables, 'numbaby' and 'numtailb', were used in a hierarchical algorithm to determine the number of births per episode. 'numtailb', if not missing, was used. Where 'numtailb' was missing, 'numbaby' (where not missing) was used. Where 'numtailb' was missing and

'numbaby' was not missing but had a filled value of "X" (unknown), the number of births in the episode was assumed to be nine.

We are aware that patid and epikey do not always uniquely identify an episode of care in the Maternity data, that is, for some patid-epikey combinations there may be multiple records with different data for the same episode of care. This may result from multiple birth pregnancies, recording maternal or birth history information during an episode of care or other anomalies in data recording. Depending on the requirements of the study, it may be necessary to exclude patients with multiple record combinations from the study or to exclude the affected records only. In some cases, multiple patid-epikey combinations may not present an issue to the study e.g. when using maternity data to determine pregnancy (yes/no).

Changes introduced in HES APC sets

Set 12

Licensing obligations require that no attempts are made to re-identify patients in CPRD datasets. The epikey variable has been encoded by CPRD to minimise the risk of breaching licensing conditions through linkage of these data to other HES data sources containing patient identifiable information. What this means is that from set 12, the epikey variable is different from that of previous sets and will differ in each future release of HES APC linkage sets.

Values of the variables 'duration', 'epidur' and 'acpdur' are now provided as generated by HES. These were previously recalculated by CPRD by adding one day to all durations where hospital admission and discharge occurred on the same day. Values are being retained as generated by HES to provide users with greater flexibility in analysis.

Set 13

The definition of the derived ethnicity variable in the patient file has been changed so that ethnicity is specified where at least one episode has a specific ethnicity recorded but the majority of values are "unknown". This is the second recent change to the ethnicity data provided. Since set 11, the original ethnicity value for each episode has been included in the hospital episodes file (hes_episodes.txt), and derived patient ethnicity (gen_ethnicity) data is based on data recorded in HES Outpatient and HES A&E data in addition to HES Admitted Patient Care.

Set 15

Health Resource Group file (hes_hrg.txt): the hrglate variable has been retired and is no longer supplied by NHS Digital. It has therefore been removed from the set 15 dataset. NHS Digital have updated the hrglate35 variable, and this data is now complete and available for the HES years 2003-2011.

Known issues

During the development process, we conferred with the HES team regarding some issues identified, in small numbers, in the data. These known issues include:

- Invalid/missing dates depicted as 15/10/1582 or 01/01/1600
- Episodes where admission date precedes the epistart date
- Unfinished episodes
- Explicit duplicate records which vary only by unique episode identifier (epikey)
- Maternity records may have inconsistencies which need to be considered when using the data
- Provisional HES data are monthly publications of HES data. These data may be incomplete or contain errors for which no adjustments have yet been made by HES. Counts produced from provisional data are likely to be lower than those generated for the same period in the final dataset.

It is also probable that clinical data are not complete, which may affect the last two months of any given period. There may also be errors due to coding inconsistencies that have not yet been investigated and corrected. At the end of the fiscal year ("month 13"), the annual data is refreshed and known data quality issues are corrected, prior to locking the annual published data.

Look-up files

CPRD do not provide ICD-10 or OPCS dictionaries.

The ICD-10 codes have been slightly modified from those provided by the WHO. We recommend acquiring lookup tables for ICD-10 codes from the NHS Digital Clinical Classifications Service by emailing them at information.standards@nhs.net or by telephoning 0300 303 4777. Note that a license is required.

It is likely that the lookup table that will be of most use to you is the ICD-10 Metadata file. This file contains all valid ICD codes, their titles, and category titles. You will be able to find out further information, including details of the licence you will need to obtain at:
https://digital.nhs.uk/article/1117/Clinical-Classifications

The Office of Population Censuses and Surveys (OPCS) Classification of Interventions and Procedures codes are also available from the NHS Digital Clinical Classifications Service:
https://digital.nhs.uk/article/1117/Clinical-Classifications

As with ICD10 codes, a license may be required to access OPCS data.

Future plans

Additional administrative years of HES data will be incorporated as they become available. Additional practices will also be added as they consent to the linkage.

Plans have been made to include data from Scotland, Wales and Northern Ireland, but there is no timescale set on when this might happen. The HES-CPRD link is part of the total linkage programme that will enable more comprehensive anonymised longitudinal patient journeys to be tracked.