Medicines & Healthcare products Regulatory Agency

CPRD

Hospital Episode Statistics (HES)

Accident & Emergency and

CPRD primary care data

Documentation (set 19)

Version 1.7

Date: 01 December 2020

NIHR | National Institute for Health Research

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	Tarita Murray-Thomas	Wilhelmine Meeraus, Arlene Gallagher
1.1	Modified	Tarita Murray-Thomas	Shivani Padmanabhan
1.2	Modified	Jenny Campbell	Rebecca Ghosh
1.3	Modified	Jenny Campbell	Elizabeth Crellin
1.4	Modified	Elizabeth Crellin	Jenny Campbell
1.5	Modified	Elizabeth Crellin	Jenny Campbell
1.6	Modified	Elizabeth Crellin	Jenny Campbell
1.7	Modified	Susan Hodgson	Jenny Campbell

Summary of Changes

Version 1.1
- Updated document version number, date and HES set
- Added the HES AE coverage dates for this release
- Added explanation of changed definition of the derived ethnicity variable
- Added changes introduced in set 13
- Updated references to reflect change of name from HSCIC to NHS Digital

Version 1.2
- Updated document version number, date and HES set for the release of set 14.
- Updated match rank table for set 14

Version 1.3
- Updated document version number, date and HES set for the release of set 15.
- Updated match rank table for set 15
- waitdays variable is no longer available and has been removed from the documentation.

Version 1.4
- Updated document version number, date and HES set for the release of set 16.
- Updated match rank table for set 16
- Updated to include CPRD Aurum
- diagscheme variable has been added

Version 1.5
- Updated document version number, date and HES set for the release of set 17.
- Updated match rank table for set 17
- Added table showing percentage capture of national A&E attendances by year
- Corrected data type for four variables in Attendance file

Version 1.6
- Updated document version number, date and HES set for the release of set 18

- Updated match rank table for set 18

Version 1.7

- Updated document version number, date and HES set for the release of set 19
- Updated match rank table for set 19
- Added sections on impact of COVID-19 pandemic and the introduction of the Emergency Care Dataset on HES A&E reporting.

HES Accident & Emergency (A&E) data linked to CPRD primary care data

This document provides an overview of HES Accident and Emergency (HES A&E) data, and the available subset that is linked to CPRD GOLD and CPRD Aurum.

Impact of the COVID-19 pandemic

Users should note there was a reduction in attendances at Emergency Departments from March 2020, at the start of the COVID-19 outbreak (see: https://digital.nhs.uk/data-and-information/publications/statistical/hospital-accident--emergency-activity/2019-20/data-quality-statement).

What are the HES Accident and Emergency data?

HES A&E data consist of individual records of patient care administered in the accident and emergency setting in England. These data are a subset of national A&E data collected by NHS England to monitor the national standard that 95% of patients attending A&E for care should wait no longer than 4 hours from arrival to admission, transfer or discharge. A&E data are submitted by providers of all types of A&E services in England - Type 1, Type 2, Type 3, Type 4 departments and urgent care centres that average more than 50 attendances per week **(Annex A).**

The collection of HES A&E was first started in April 2007 on an experimental basis and continued until 2012/2013 when the experimental label was lifted. Table 1 shows the percentages of national A&E attendances that were captured in HES A&E data by year, excluding planned follow-up attendances:

Table 1: Capture of HES A&E attendances

Year	Experimental status	Percentage attendances captured
2007/08	Experimental	62%
2008/09	Experimental	68%
2009/10	Experimental	74%
2010/11	Experimental	74%
2011/12	Experimental	80%
2012/13		83%
2013/14		83%
2014/15		86%

Records in the HES A&E database are called ‘attendances’ and each A&E attendance relates to a single visit by an individual to A&E. Where follow up care is required and provided by the A&E department, a second planned attendance is recorded. A&E data collected includes details about patients’ attendance, outcome of the visit to A&E, waiting times, referral source, A&E diagnosis (not ICD-10 coded), A&E treatment (drug prescribing not recorded), A&E investigations (not OPCS coded) and Health Resource Group.

HES A&E may be used to further clarify the health care pathway, to quantity health resource use and costs in the emergency setting, and to assess variations in the uptake of emergency services over time. Before requesting HES A&E data, users are encouraged to familiarise themselves with the content of HES A&E data. Details on the fields available 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 A&E 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

The impact of the introduction of the Emergency Care Data Set on HES A&E reporting

Overall coverage in HES has increased from 2018-19, however the data completeness for a number of key fields has reduced since the phased introduction of the new Emergency Care Data Set (ECDS) commenced in October 2017. As such, Annual data for certain fields and reported activity is now no longer directly comparable.

More information of the impact of the introduction of the ECDS on HES A&E is available from: https://digital.nhs.uk/data-and-information/publications/statistical/hospital-accident--emergency-activity/2019-20/data-quality-statement

The methodological change notice paper published by NHS Digital is available at: https://digital.nhs.uk/binaries/content/assets/website-assets/publications/publications-admin-pages/methodological-changes/methchange20171212_hes.pdf

Accessing HES Accident and Emergency data linked to CPRD GOLD and CPRD Aurum

HES A&E 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 the CPRD for a fee subject to MHRA Independent Scientific Advisory Committee (ISAC) 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 to have linked HES data.

Linkage coverage period

The latest release of HES A&E data linked to CPRD primary care data (set 19) covers the period **April 2007 – March 2020**. Data up to March 2019 are final. Data for 2019/2020 (April 2019 – March 2020) are provisional.

Linkage algorithm and the match_rank variable

Linkage between HES A&E and CPRD primary care data uses an eight-step deterministic linkage algorithm based on four identifiers, shown in Table 2. 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 A&E data.

Table 2: 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 3A and 3B.

Table 3A: Number and proportion of **CPRD GOLD** patients matched to a HES patient* at each step of the linkage algorithm in set 19.

Linkage step (match_rank)	Frequency	Percent
1	5,610,711	68.5%
2	2,296,650	28.0%
3	13,339	0.2%
4	17,927	0.2%
5	3,424	0.0%
6	233,506	2.8%
7	14,344	0.2%
8	6,410	0.1%

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

Table 3B: Number and proportion of **CPRD Aurum** patients matched to a HES patient* at each step of the linkage algorithm in set 19.

Linkage step (match_rank)	Frequency	Percent
1	19,979,123	65.6%
2	9,299,237	30.5%
3	41,808	0.1%
4	64,613	0.2%
5	10,770	0.0%
6	978,296	3.2%
7	55,875	0.2%
8	26,313	0.1%

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

CPRD provides users with a match_rank variable which corresponds to the step at which the 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 A&E 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 to the ISAC.

A linkage coverage file (linkage_coverage.txt) provides the start and end dates of HES encounter time.

A minority of patients are linked to multiple HESIDs. These patients are removed from the HES A&E dataset. However, the data have been retained 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 to the ISAC.

As far as possible, the linked HES A&E 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.

Data structure and formatting

HES A&E data provided by the CPRD represents only a subset of the variables that are collected in the National HES A&E dataset provided by NHS Digital. Fields such as organisation fields which may lead to the potential re-identification of patients or practices are not collected by the CPRD and/or not supplied to users.

The data are arranged into files relating to A&E attendance, diagnosis, treatment, and investigations undertaken in the A&E setting. Each record represents a single A&E attendance at a single provider. The HES 'aekey', the record identifier created by HES, is unique in combination with the CPRD patient identifier (**patid**). The patient identifier (**patid**) may be used to link together A&E attendance records for a single patient with CPRD HES admitted patient care and/or outpatient records.

For each patient cohort, HES A&E data will be provided as separate text tab delimited files. 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:

- CPRD has introduced the HES patient identifier (**gen_hesid**) in A&E data. This is unique across all CPRD linked HES datasets including HES admitted patient care (APC), HES Outpatient (OP) and HES A&E data. An individual that has contributed data to more than one CPRD practice will have the same gen_hesid but this may change between linkage sets.
- The aekey variable has been altered so that it is unique (by patient identifier) across all A&E data.
- CPRD has provided a derived ethnicity variable (**gen_ethnicity**) in the HES A&E patient file which is the most commonly recorded ethnicity for each patient, among all HES data including HES APC, HES OP and HES A&E. The ethnicity recorded at A&E attendance (**ethnos**) remains unchanged.

Changes introduced in HES A&E sets

Set 12

Licensing obligations require that no attempts are made to re-identify patients in CPRD datasets. The aekey has been encoded by the 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 aekey variable is different from that of previous sets and will differ in each future release of HES A&E linkage sets.

Set 13
The definition of the derived ethnicity variable (gen_ethnicity) 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”.

Set 16
The diagscheme variable has been added to the CPRD linked dataset.

Known issues

- Data coverage is incomplete in comparison to national A&E data attendances. Analyses of patterns of missing data may be required when using these data.
- Some variables have high levels of missing data e.g. duration to initial assessment (initdur)
- Provisional HES A&E 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 there is a “month 13” annual refresh which corrects known data quality issues prior to locking the annual published data.
- Diagnosis is recorded using a specific coding system for HES A&E data. However, there may be some ICD-10 and Read codes within the A&E data resulting from experimental recording in a small percentage of A&E departments. The diagscheme variable provides an indication of the coding scheme used.

Look-up files
Lookup files relating to the use of HES A&E data will not be provided by the CPRD. These can be obtained online from NHS Digital using this link: https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/hospital-episode-statistics/hospital-episode-statistics-data-dictionary

Annex A: Type of A&E service providers

A&E Service type	*Description*
Type 1	A&E department = A consultant led 24-hour service with full resuscitation facilities and designated accommodation for the reception of accident and emergency patients
Type 2	A&E department = A consultant led single specialty accident and emergency service (e.g. ophthalmology, dental) with designated accommodation for the reception of patients
Type 3 / Type 4 / Urgent Care Centre	Other type of A&E/minor injury units (MIUs)/Walk-in Centres (WiCs)/Urgent Care Centre primarily designed for the receiving of accident and emergency patients. A type 3 department may be doctor led or nurse led. It may be co-located with a major A&E or sited in the community. A defining characteristic of a service qualifying as a type 3 department is that it treats at least minor injuries and illnesses (sprains for example) and can be routinely accessed without appointment.

HES A&E: Data dictionary

1. Patient (hesae_patient.txt)

Column name	*Description*	*Type*	*Format*
patid	Encrypted unique key given to a patient in CPRD GOLD or CPRD Aurum [primary key]	INTEGER	20
pracid	Encrypted unique key given to a practice in CPRD GOLD or CPRD Aurum	INTEGER	5
gen_hesid[1]	A generated unique key assigned to a patient across all CPRD linked HES datasets within a linkage set. An individual that has contributed data to more than one CPRD practice has the same gen_hesid but this may change between linkage sets.	INTEGER	20
n_patid_hes[1]	Number of individuals in CPRD GOLD or CPRD Aurum assigned the same gen_hesid (unique patient identifier generated in HES)	INTEGER	3
gen_ethnicity[1]	Patient's ethnicity derived from all HES data (including HES outpatient, HES admitted patient care and HES A&E)	CHAR	10
match_rank[2]	Indicates the quality of matching between a record in HES and CPRD primary care data and gives the level of confidence that an HES record has been correctly matched to a patient in CPRD GOLD or CPRD Aurum.	INTEGER	1

[1] Variable generated by CPRD.

[2] An eight-step process is used to match patients in CPRD primary care data (CPRD GOLD or CPRD Aurum) and HES using some or all of the following: NHS number, date of birth, sex and postcode. Only data for patients matched using steps 1-5 has been provided.

2. ATTENDANCE (hesae_attendance.txt)

Column name	*Description*	*Type*	*Format*
patid	Encrypted unique key given to a patient in CPRD GOLD or CPRD Aurum [primary key, in combination with aekey]	INTEGER	20
aekey[3]	Record identifier (unique in combination with patid) [primary key, in combination with patid]	INTEGER	20
arrivaldate	The arrival date of a patient in the A&E department	DATE	dd/mm/yyyy
aepatgroup	The reason for an A&E episode	INTEGER	2
aeattendcat	An indication of whether a patient is making an initial or follow-up attendance within a particular A&E department	INTEGER	1
aearrivalmode	The mode by which a patient arrived at an A&E department	INTEGER	1
aedepttype	A classification of A&E department type according to the activity carried out	INTEGER	2
aerefsource	The source of referral for each A&E episode	INTEGER	2
aeincloctype	Classification of the place where the incident occurred that led to an A&E episode	INTEGER	2
aeattenddisp	The way in which an A&E attendance might end	INTEGER	2
initdur	The time (expressed as a whole number of minutes) between the patient's arrival and their initial assessment	INTEGER	8
tretdur	The time (expressed as a whole number of minutes) between the patient's arrival and the start of their treatment	INTEGER	8
concldur	The time (expressed as a whole number of minutes) between the patient's arrival and conclusion of their attendance or treatment (whichever is later)	INTEGER	8
depdur	The time (expressed as a whole number of minutes) between the patient's arrival, and the time the A&E attendance has concluded, and the department is no longer responsible for the care of the patient	INTEGER	8
ethnos	Ethnic category recorded at attendance	CHAR	2

[3] This variable has been altered by the CPRD so that it is unique within and across all HES years.

3. DIAGNOSIS (hesae_diagnosis.txt)

Column name	*Description*	*Type*	*Format*
patid	Encrypted unique key given to a patient in CPRD GOLD or CPRD Aurum [primary key, in combination with aekey]	INTEGER	20
aekey[3]	Record identifier (unique in combination with patid) [primary key, in combination with patid]	INTEGER	20
diag	A&E diagnosis - 6 characters. A 6-character code made up of diagnosis condition (n2), sub-analysis (n1), anatomical area (n2) and anatomical side (an1). Only certain diagnoses contain a sub-analysis	CHAR	6
diag2	A&E diagnosis - 2 characters. Includes the diagnosis condition (n2) of the 6-character diagnosis code	CHAR	2
diag3	A&E diagnosis - 3 characters. Includes diagnosis (n2) and the sub-analysis (n1) of the 6-character diagnosis code. If no sub-analysis has been provided, or is not applicable, then the 2-character description is displayed if available.	CHAR	3
diaga	A&E diagnosis - anatomical area	CHAR	2
diags	A&E diagnosis - anatomical side	CHAR	1
diagscheme	Coding scheme in use	INTEGER	1
diag_order[1]	Ordering of diagnosis at attendance, within range 1-12	INTEGER	2

4. INVESTIGATION (hesae_investigation.txt)

Column name	*Description*	*Type*	*Format*
patid	Encrypted unique key given to a patient in CPRD GOLD or CPRD Aurum [primary key, in combination with aekey]	INTEGER	20
aekey[3]	Record identifier (unique in combination with patid) [primary key, in combination with patid]	INTEGER	20
invest	A&E investigation - 6 characters. A 6-character code made up of investigation (n2) and local sub-analysis (up to an4)	CHAR	6
invest2	A&E investigation - 2 characters. Consists of the investigation (n2) of the 6-character investigation code	CHAR	2
invest_order[1]	Ordering of investigation at attendance, within range 1-12	INTEGER	2

[1] Variable generated by CPRD.
[3] This variable has been altered by the CPRD so that it is unique within and across all HES years.

5. TREATMENT (hesae_treatment.txt)

Column name	*Description*	*Type*	*Format*
patid	Encrypted unique key given to a patient in CPRD GOLD or CPRD Aurum [primary key, in combination with aekey]	INTEGER	20
aekey[3]	Record identifier (unique in combination with patid) [primary key, in combination with patid]	INTEGER	20
treat	A&E Treatment - 6 characters. Treatment code made up of treatment (n2), sub-analysis (n1) and a local use section (up to an3)	CHAR	6
treat2	A&E Treatment - 2 characters. Consists of treatment (n2) of the 6-character treatment code	CHAR	2
treat3	A&E Treatment - 3 characters. Consists of treatment (n2) and the sub-analysis (n1) of the 6-character treatment code. If no sub-analysis has been provided, or is not applicable, then the 2-character description is displayed if available	CHAR	3
treat_order[1]	Ordering of treatment at attendance, within range 1-12	INTEGER	2

6. HEALTH RESOURCE GROUP TABLE (hesae_hrg.txt)

Column name	*Description*	*Type*	*Format*
patid	Encrypted unique key given to a patient in CPRD GOLD or CPRD Aurum [primary key, in combination with aekey]	INTEGER	20
aekey[3]	Record identifier (unique in combination with patid) [primary key, in combination with patid]	INTEGER	20
domproc	Dominant Procedure	CHAR	6
hrgnhs	Trust derived HRG value	CHAR	3
hrgnhsvn	Version number of trust derived HRG	CHAR	3
sushrg	The SUS[4] Payment by Result (PbR) derived healthcare resource group (HRG) code	CHAR	6
sushrgvers	SUS[4] generated HRG version number	NUMERIC	4

7. PATIENT PATHWAY (hesae_pathway.txt)

Column name	*Description*	*Type*	*Format*
patid	Encrypted unique key given to a patient in CPRD GOLD or CPRD Aurum [primary key, in combination with aekey]	INTEGER	20
aekey[3]	Record identifier (unique in combination with patid) [primary key, in combination with patid]	INTEGER	20
rttperstart[5]	The start date, for the referral to treatment period	DATE	dd/mm/yyyy
Rttperend[5]	The end date, for the referral to treatment period	DATE	dd/mm/yyyy

[1] Variable generated by CPRD
[3] This variable has been altered by the CPRD so that it is unique within and across all HES years
[4] Secondary User Services
[5] In set 19, for the 2019/20 data, we observe that a proportion of the 'rttperstart' and 'rttperend' variables have a value '1900-01-01'. We are advised that some Patient Administration Systems default to this date when no date is entered, although this is not an official default value, and should be treated as invalid.