Medicines & Healthcare products
Regulatory Agency

CPRD

Hospital Episode Statistics (HES) Outpatient Care and CPRD primary care data Documentation (set 21)

Version: 2.0

Date: 12 August 2021

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	Tarita Murray-Thomas	Helen Strongman
1.1	Formatted	Grant Lee	Wilhelmine Meeraus
1.2	Modified	Wilhelmine Meeraus	Tarita Murray-Thomas
1.3	Modified	Wilhelmine Meeraus	Tarita Murray-Thomas, Arlene Gallagher
1.4	Modified	Shivani Padmanabhan	Tarita Murray-Thomas
1.5	Modified	Tarita Murray-Thomas	Shivani Padmanabhan
1.6	Modified	Tarita Murray-Thomas	Sonia Coton
1.7	Modified	Sonia Coton	Tarita Murray-Thomas
1.8	Modified	Tarita Murray-Thomas	Susan Hodgson
1.9	Modified	Tarita Murray-Thomas	Sonia Coton, Rebecca Ghosh
2.0	Modified	Helen Booth	Mia Harley

Summary of Changes

Version 1.1
- Formatted with new agency branding and updated document title
- Included version of HES on front page

Version 1.2
- Updated document version number, date, HES set, and end of the HES OP coverage date
- Clarified information relating to the ‘match_rank’ variable under ‘HES Outpatient data and GOLD’
- Included new section under ‘Data structure and formatting’ to record changes introduced in set 11
- Updated the information relating to ethnicity under ‘Known Issues’.

Version 1.3
- Updated document version number, date, HES set, and end of the HES OP coverage date
- Added information on the coverage period of HES OP data by HES
- 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 changes introduced in set 12
- Added information under ‘Known issues’ relating to provisional HES data
- Added the data dictionary to the end of the document

Version 1.4
- Updated document version number, date, HES set, and end of the HES OP coverage date
- 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 1.5
- Updated document version number, date, HES set, and end of the HES OP coverage date

Version 1.6

- Updated document version number, date, HES set, and end of the HES OP coverage date

Version 1.7
- Updated document version number, date, HES set and end of the HES OP coverage date
- Updated to include CPRD Aurum

Version 1.8
- Updated document version number, date, HES set and end of the HES OP coverage date

Version 1.9
- Updated document version number, date, HES set and end of the HES OP coverage date
- Updated NIHR logo
- Specified primary keys in each data table

Version 2.0
- Updated document version number, date, HES set and end of the HES OP coverage date; added DOIs

HES Outpatient Care (OP) data linked to CPRD primary care data

This document provides an overview of HES Outpatient (HES OP) data, and the available subset that is linked to CPRD GOLD and CPRD Aurum.

What are the HES Outpatient data?

HES OP data are a collection of individual records of outpatient appointments occurring in England only. It includes information on the type of outpatient consultation appointment dates, the main specialty and treatment specialty under which the patient was treated, referral source, waiting times, clinical diagnosis and procedures performed.

HES OP data were collected by the NHS for the first time in 2003-04. NHS Digital (formerly known as the Health and Social Care Information Centre) first released HES OP data on an 'experimental' basis at the end of July 2006. Since then, the experimental label has been removed and in 2008 the outpatient published tables were accredited as a National Statistic. There may be potential data quality issues in the early periods when the data were collected experimentally.

HES OP data can be used to support health resource utilisation studies, clarify clinical health care pathways, and enable variations in the uptake of services to be evaluated, for example by gender and age. Before requesting HES OP data, users are encouraged to familiarise themselves with the content of HES OP 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 Outpatient 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

Accessing HES Outpatient Care data linked to CPRD GOLD and CPRD Aurum

HES OP 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 for a fee 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 to have linked HES data.

Linkage coverage period

The latest release of HES OP data linked to CPRD primary care data (set 21) covers the period **April 2003 – October 2020**. Please note that the data for 2020 (April 2020 – October 2020) are provisional HES data, up to Month 7.

DOI

Please cite in any publications using these data:

CPRD GOLD HES OP August 2021 - https://doi.org/10.48329/cp5e-7790

CPRD Aurum HES OP August 2021 - https://doi.org/10.48329/7hm3-gt75

Linkage algorithm and the match_rank variable

Linkage between HES OP 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 OP 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 21.

Linkage step (match_rank)	Frequency	Percent
1	5,679,805	68.6%
2	2,310,259	27.9%
3	13,415	0.2%
4	18,075	0.2%
5	3,447	0.0%
6	234,817	2.8%
7	14,425	0.2%
8	6,494	0.1%

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

Table 2B: Number and proportion of **CPRD Aurum** patients matched to a HES patient* at each step of the linkage algorithm in set 21.

Linkage step (match_rank)	Frequency	Percent
1	23,508,210	65.9%
2	10,772,078	30.2%
3	49,245	0.1%
4	75,154	0.2%
5	13,358	0.0%
6	1,129,843	3.2%
7	65,212	0.2%
8	32,727	0.1%

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

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 OP 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.

Modified linkage eligibility files are available upon request for records matched in steps 6-8 and for records linked to multiple HESIDs (see below). 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 OP 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.

As far as possible, the linked HES OP data is supplied “as is” (e.g. a value of “X” [from 2008/09] or “9” [prior to 2008/09] represents an unknown in an otherwise numeric field, and values of “99” or “&” represent unknowns in an otherwise string field), 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 OP data provided by CPRD represents only a subset of the variables that are collected in the National HES OP 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 appointments and contact in the outpatient clinic setting. Patients may see the same consultant, or members of his/her team, on more than one occasion during a single year. They may attend more than one consultant clinic in the same or different specialties, or in different providers, for the same or different conditions. Each record represents a single outpatient attendance at a consultant or allied health clinic at a single hospital provider. The patient identifier (**patid**) may be used to link together outpatient attendance records for a single patient with CPRD HES admitted patient care and/or A&E attendance records.

For each patient cohort, HES OP 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.

Changes introduced in HES OP sets

set 11

- CPRD introduced the HES patient identifier (**gen_hesid**) in OP data. This is unique across all CPRD linked HES datasets including HES admitted patient care (APC), HES OP and HES accident and emergency (A&E) data.
- The attendkey variable has been altered so that this is now unique (by patient identifier) across all HES OP data.
- CPRD provided a derived ethnicity variable (**gen_ethnicity**) 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 the outpatient appointment (**ethnos**) remains unchanged.

set 12

Licensing obligations require that no attempts are made to re-identify patients in CPRD datasets. The attendkey variable 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 attendkey variable is different from that of previous sets and will differ in each future release of HES OP 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”.

Known issues

Some fields, which are of great potential interest, were not designated as mandatory when the dataset was originally designed. This has led to apparently low coding levels over the years. These fields include:

- **Ethnic Group** (low levels of coding): Recording of ethnicity is lower in the earlier years of collection and better in the later years.
- **Diagnosis** (less than 5% of all attendances): Unlike HES admitted patient care data, it is not mandatory to record diagnostic information in HES OP using ICD-10 codes. For this reason, diagnostic information is not widely available in this data offering. HES OP data will therefore have limited use in epidemiology research where clinical diagnostic data is needed.
- **Procedure** (about 5-15% of all attendances): Procedures undertaken in the outpatient setting can be identified in HES OP data. However, the completeness of recording of these data for select procedures is unknown.
- There are minor problems with national data coverage reported with each annual dataset.

Considering the limitations listed above, a conservative approach would be to use HES OP data from 2007 onwards.

Provisional HES OP 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.

Look-up files

Lookup files relating to the use of HES OP 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

HES OP: Data dictionary

1. Patient (hesop_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. This variable was first available with HES set 10.

2. System/Patient pathway (hesop_patient_pathway_YYYY.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 attendkey]	INTEGER	20
attendkey[3]	Record identifier (unique in combination with patid) [primary key, in combination with patid]	INTEGER	20
perend	RTT[4] period end date	DATE	dd/mm/yyyy
perstart	RTT[4] period start date	DATE	dd/mm/yyyy
subdate	Submission date	DATE	dd/mm/yyyy
HES_yr[1]	Events recorded between 01/04/YYYY- 31/03/YYYY+1 inclusive. e.g. events with a HES_yr of 2006 would be dated between 01/04/2006 - 31/03/2007, inclusive	INTEGER	4

[1] Variable generated by CPRD.

[3] This variable has been altered so it is now unique (by patid) within and across all HES years.

[4] RTT: Referral To Treatment

3. Appointment (hesop_appointment_YYYY.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 attendkey]	INTEGER	20
attendkey[3]	Record identifier (unique in combination with patid) [primary key, in combination with patid]	INTEGER	20
ethnos	Ethnic category as recorded at appointment	CHAR	11
admincat	Administrative category 1=NHS, 2=Private	INTEGER	8
apptdate	Appointment date	DATE	dd/mm/yyyy
apptage	Age on day of appointment	INTEGER	8
atentype	Attendance type	INTEGER	8
attended	Attended or did not attend	INTEGER	8
dnadate	Last DNA or patient cancelled date	DATE	dd/mm/yyyy
firstatt	First attendance	CHAR	1
outcome	Outcome of attendance	INTEGER	8
priority	Priority type	INTEGER	8
refsourc	Source of referral	INTEGER	8
reqdate	Referral request received date	DATE	dd/mm/yyyy
servtype	Service type requested	INTEGER	8
stafftyp	Medical staff type seeing patient	INTEGER	8
wait_ind	Waiting calculation indicator	INTEGER	8
waiting	Days waiting	INTEGER	8
HES_yr[1]	Events recorded between 01/04/YYYY- 31/03/YYYY+1 inclusive. e.g. events with a HES_yr of 2006 would be dated between 01/04/2006 - 31/03/2007, inclusive	INTEGER	4

[1] Variable generated by CPRD.

[3] This variable has been altered so it is now unique (by patid) within and across all HES years.

4. Clinical (hesop_clinical_YYYY.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 attendkey]	INTEGER	20
attendkey[3]	Record identifier (unique in combination with patid) [primary key, in combination with patid]	INTEGER	20
diag_01	Primary diagnosis	CHAR	8
diag_02	Secondary diagnosis	CHAR	8
diag_03[5]	Secondary diagnosis	CHAR	8
diag_04[5]	Secondary diagnosis	CHAR	8
diag_05[5]	Secondary diagnosis	CHAR	8
diag_06[5]	Secondary diagnosis	CHAR	8
diag_07[5]	Secondary diagnosis	CHAR	8
diag_08[5]	Secondary diagnosis	CHAR	8
diag_09[5]	Secondary diagnosis	CHAR	8
diag_10[5]	Secondary diagnosis	CHAR	8
diag_11[5]	Secondary diagnosis	CHAR	8
diag_12[5]	Secondary diagnosis	CHAR	8
opertn_01	Main (i.e. most resource intensive) operation	CHAR	8
opertn_02	Secondary operation/procedure	CHAR	8
opertn_03	Secondary operation/procedure	CHAR	8
opertn_04	Secondary operation/procedure	CHAR	8
opertn_05	Secondary operation/procedure	CHAR	8
opertn_06	Secondary operation/procedure	CHAR	8
opertn_07	Secondary operation/procedure	CHAR	8
opertn_08	Secondary operation/procedure	CHAR	8
opertn_09	Secondary operation/procedure	CHAR	8
opertn_10	Secondary operation/procedure	CHAR	8
opertn_11	Secondary operation/procedure	CHAR	8
opertn_12	Secondary operation/procedure	CHAR	8
opertn_13[6]	Secondary operation/procedure	CHAR	8
opertn_14[6]	Secondary operation/procedure	CHAR	8
opertn_15[6]	Secondary operation/procedure	CHAR	8
opertn_16[6]	Secondary operation/procedure	CHAR	8
opertn_17[6]	Secondary operation/procedure	CHAR	8
opertn_18[6]	Secondary operation/procedure	CHAR	8
opertn_19[6]	Secondary operation/procedure	CHAR	8
opertn_20[6]	Secondary operation/procedure	CHAR	8
opertn_21[6]	Secondary operation/procedure	CHAR	8
opertn_22[6]	Secondary operation/procedure	CHAR	8
opertn_23[6]	Secondary operation/procedure	CHAR	8
opertn_24[6]	Secondary operation/procedure	CHAR	8
operstat	Operation status code	CHAR	8
tretspef	Treatment speciality	CHAR	8
mainspef	Main speciality	CHAR	8

[5] Only diag_01 and diag_02 were available prior to 01/04/2007.
[6] Only 12 operation status codes were available prior to 01/04/2007.

HES_yr[1]	Events recorded between 01/04/YYYY- 31/03/YYYY+1 inclusive. e.g. events with a HES_yr of 2006 would be dated between 01/04/2006 - 31/03/2007, inclusive	INTEGER	4

[1] Variable generated by CPRD.