Medicines & Healthcare products Regulatory Agency

CPRD

Hospital Episode Statistics (HES) Diagnostic Imaging Dataset and CPRD primary care data Documentation (Set 17)

Version: 1.5

Date: 2 April 2019

NHS 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	Jenny Campbell	Arlene Gallagher Shivani Padmanabhan
1.1	Modified	Jenny Campbell	Shivani Padmanabhan
1.2	Modified	Jenny Campbell	Rebecca Ghosh
1.3	Modified	Jenny Campbell	Rebecca Ghosh, Arlene Gallagher
1.4	Modified	Jenny Campbell	Jessie Oyinlola
1.5	Modified	Jenny Campbell	

Summary of Changes

Version 1.1

- Updated document version number, date and HES Set
- Added the HES DID coverage dates for 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
- Added the HES DID coverage dates for set 14

Version 1.3

- Updated document version number, date and HES Set
- Added the HES DID coverage dates for set 15
- Updated header and footer with new agency branding

Version 1.4

- Updated document version number, date and HES Set
- Added the HES DID coverage dates for set 16
- Updated to include CPRD Aurum

Version 1.5

- Updated document version number, date and HES Set
- Added the HES DID coverage dates for set 17

Medicines & Healthcare products Regulatory Agency

CPRD

HES Diagnostic Imaging Dataset (DID) and CPRD primary care data

This document provides an overview of the HES Diagnostic Imaging Dataset (HES DID) data, and the available subset that is linked to CPRD GOLD and CPRD Aurum.

What are the HES Diagnostic Imaging Dataset data?

The Diagnostic Imaging Dataset (DID) are a collection of detailed information about diagnostic imaging tests carried out on NHS patients in England. The DID includes information on imaging tests carried out from 1 April 2012. It does not include the images that are produced as a result of these tests. The DID data supplied captures information about referral source and patient type, details of the test (type of test and body site), plus items about waiting times for each diagnostic imaging event, from time of test request through to time of reporting. The DID enables analysis of demographic and geographic variation in access to different test types and different providers.

The DID data are routinely linked to Hospital Episode Statistics (HES) through NHS Digital, formerly known as the Health and Social Care Information Centre to create the HES DID dataset. A stepwise deterministic linkage algorithm is used based on four identifiers as shown in the table below. The matching steps are applied sequentially. If a DID record is matched in one step it is no longer available for matching in subsequent steps. As records are matched against more criteria in the earlier stages, these records may be regarded as having the strongest match. Matching at Stage 1A therefore provides the greatest level of confidence that a DID record has been correctly matched to a patient in HES.

Step	NHS Number	Date of Birth (DOB)	Gender	Postcode	Notes
1A	Exact	Exact	Exact	Exact	
1B	Exact	Exact		Exact	
2	Exact	Exact	Exact		
3	Exact	Partial	Exact	Exact	
4	Exact	Partial	Exact		
5	Exact			Exact	
6		Exact	Exact	Exact	Where NHS number does not contradict the match and DOB is not 1 January and the Postcode is not in the 'ignore' list
7		Exact	Exact	Exact	Where NHS number does not contradict the match and DOB is not 1 January
8	Exact				

Linkage of HES and DID data is performed at the record level. The match rank variable (hes_did_matchrank) corresponding to the step at which the match was established between a DID record and a HES patient can be found in the DID Referral file (hesdid_referral.txt).

NHS National Institute for Health Research

Medicines & Healthcare products Regulatory Agency

CPRD

Accessing HES Diagnostic Imaging Dataset linked to CPRD GOLD and CPRD Aurum

HES DID 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. Please contact CPRD Enquires regarding the cost for access.

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

Linkage coverage period

The latest release of HES DID data linked to CPRD primary care data (set 17) covers the period **April 2012 – November 2018**. Please note that the data for 2018/2019 (April 2018 – November 2018) are provisional HES data, up to Month 8.

Linkage algorithm and the match_rank variable

Linkage between HES DID 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 DID data.

Table 1: NHS Digital 8 step linkage algorithm

Step	Match
1	Exact NHS number, sex, 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 the NHS number does not contradict the match, the DoB is not 1st of January and the postcode is 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.

NHS National Institute for Health Research

Table 2A: Number and proportion of **CPRD GOLD** patients matched to a HES patient* at each step of the linkage algorithm in set 17.

Linkage step (match rank)	Frequency	Percent
1	5322358	67.98
2	2221441	28.37
3	13192	0.17
4	17506	0.22
5	3448	0.04
6	230562	2.95
7	14047	0.18
8	6318	0.08

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

Table 2B: Number and proportion of **CPRD Aurum** patients matched to a HES patient* at each step of the linkage algorithm in set 17.

Linkage step (match rank)	Frequency	Percent
1	12139425	65.13
2	5755336	30.88
3	26478	0.14
4	40291	0.22
5	6347	0.03
6	618657	3.32
7	36732	0.20
8	15250	0.08

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

Linkage of CPRD and HES DID data is performed at the patient level. CPRD provides users with a match_rank variable in the DID Patient file (hesdid_patient.txt) which corresponds to the step at which the match between a patient in CPRD and HES DID 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 DID dataset.

As far as possible, the linked HES DID 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

Data structure and formatting

HES DID data provided by CPRD represents only a subset of the variables that are collected in the National HES DID 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 CPRD and/or not supplied to users.

For each patient cohort, HES DID data will be provided as separate text tab delimited files. The data are arranged into files relating to patient information, referral information and information about the imaging test which was done. 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.

Licensing obligations require that no attempts are made to re-identify patients in CPRD datasets. The 'submissiondataid' 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. This means that the 'submissiondataid' variable will differ in each release of HES DID linkage sets.

Known issues

- Ethnic Group: Patients have an ethnic category recorded for each referral and this varies.
- Provisional HES data: Provisional HES DID data are monthly publications of 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.
- Duplicate submissiondataid: During processing, we identified a small number of records which were not unique based on 'patid' and 'submissiondataid'. These rows have matching dates and are therefore assumed to be duplicates of another submission, however; they differ in other fields such as *ic_reftype_desc*. All 'submissiondataid' have been modified from 12 to 13 digits by adding a 1, where duplicates have been assigned a 'submissiondataid' ending in 2.

Medicines & Healthcare products Regulatory Agency

CPRD

HES DID: Data dictionary

1. Patient file (hesdid_patient.txt)

Column name	*Description*	*Type*	*Format*
patid	Encrypted unique key given to a patient in CPRD GOLD or CPRD Aurum	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
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. DID Referral (hesdid_referral.txt)

Column name	*Description*	*Type*	*Format*	*Lookup (if applicable)*
patid	Encrypted unique key given to a patient in CPRD GOLD or CPRD Aurum	INTEGER	20	
submissiondataid	Record identifier (unique in combination with patid)	INTEGER	13	
did_ethcat	Ethnic code as per submission. Contains nationally defined codes. Left blank where a value was not supplied	STRING	2	ethcat_desc
ic_reftype_desc	Referrer Type Description (GP, Consultant, Nurse, Physio, Other health professional, not known)	STRING	30	
ic_prov_shacode	SHA code representing the SHA with which the provider organisation is associated	STRING	3	prov_shaname
did_patsource_code	Patient Source Setting. Categorises the type of department or organisation making the referral for imaging activity	INTEGER	2	patsource_desc
did_date1	Diagnostic Test Request Date. The date that the referrer made the referral request. Date submitted must be <= Diagnostic Test Request Received Date and >1 year before Diagnostic Test Date. Date can be left blank	DATE	dd/mm/yyyy	
did_date2	Diagnostic Test Request Received Date. The date that the diagnostic provider received the referral request. Date must be >= Diagnostic Test Request Date and >1 year before Diagnostic Test Date. Date can be left blank	DATE	dd/mm/yyyy	
hes_did_matchrank	Indicates the quality of matching between HES and DID and gives the level of confidence that a DID record has been correctly matched to a patient in HES. This match rank is calculated per DID record	STRING	2	

3. DID Test (hesdid_test.txt)

Column name	*Description*	*Type*	*Format*	*Lookup (if applicable)*
patid	Encrypted unique key given to a patient in CPRD GOLD or CPRD Aurum	INTEGER	20	
submissiondataid	Record identifier (unique in combination with patid)	INTEGER	13	
fyear	Financial year in which the imaging test was carried out	INTEGER	4	
did_date3	Diagnostic Test Date. The date that the referrer made the referral request. It must be in the correct CCYY-MM-DD format and be <= Diagnostic Test Request Received Date. Diagnostic Test Request Date can be left blank	DATE	dd/mm/yyyy	
did_date4	Service Report Issue Date. The date that the diagnostic provider issues the test report. Date must be >1 month after Diagnostic Test Date	DATE	dd/mm/yyyy	
did_nicip_code	National Interim Clinical Imaging Procedure Code (NICIP)	STRING	9	nicip_desc
did_snomedct_code	Imaging Code (SNOMED-CT). It must be a valid active code	INTEGER	12	snomedct_desc
ic_modality_id	Modality ID. Broad categories represented by an ID, grouping procedures or methods used for examination that may include procedures assisted by the method, e.g. biopsy or injection. Derived from submitted NICIP / SNOMED CT codes	INTEGER	12	modality_desc
ic_sub_modality_id	Sub modality ID. Narrower categories represented by an ID, grouping procedures or methods used for examination that may include procedures assisted by the method, e.g. biopsy or injection, which are encompassed within each modality category. These categories are derived from submitted NICIP or SNOMED-CT codes	INTEGER	12	sub_modality_desc
ic_region_id	Body region code. Broad categories represented by an ID	INTEGER	12	region_desc

	categorising submitted NICIP or SNOMED-CT by body region			
ic_sub_region_id	Body sub region code. Narrower categories represented by an ID categorising submitted NICIP or SNOMED-CT to provide a more complete categorisation of imaging by body sub region	INTEGER	12	sub_region_desc
ic_system_id	Body structure/system groups code. Broad categories represented by an ID categorising submitted NICIP or SNOMED-CT by body structure or system groups as identified by Clinical Terminologists at NHS Digital	INTEGER	12	system_desc
ic_sub_sys_id	Body sub system group code. Sub-classification within system categorisation of submitted NICIP or SNOMED-CT codes to provide a more complete categorisation by imaging by body subsystem	INTEGER	12	sub_sys_desc
ic_sub_syscomp_id	Body sub system component code. Further sub-classification within system and sub-system categorisation of submitted NICIP / SNOMED-CT codes by body structure / system groups	INTEGER	12	sub_syscomp_desc
ic_morhpology_id	Morphology ID. Based on submitted NICIP / SNOMED-CT codes. Provides categorisation of imaging conducted on morphologic abnormalities (alterations of the body structure from its original anatomical structure)	INTEGER	12	morph_desc
ic_fetal_id	Fetal ID. Based on submitted NICIP or SNOMED-CT codes. Provides categorisation of imaging on fetal body structures	INTEGER	12	fetal_desc
ic_cancer_desc	Description of category used to diagnose or discount cancer	STRING	40	
ic_sub_cancer_desc	Descriptive sub-division of each identified cancer category	STRING	45	