



# **Linkage Source Data Documentation**

**Version 1.3**

**Date: 14 November 2024**

## 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	Shivani Puri, Arlene Gallagher	Helen Strongman, Rachael Boggon
1.1	Modified	Shivani Puri	
1.2	Modified	Rachael Boggon	
1.3	Modified	Shivani Puri, Helen Strongman	Arlene Gallagher
1.4	Modified	Shivani Puri	Helen Strongman
1.5	Modified	Arlene Gallagher	Helen Strongman
1.6	Formatted	Grant Lee	
1.7	Modified	Helen Strongman	Arlene Gallagher Shivani Padmanabhan
1.8	Modified	Helen Strongman	Arlene Gallagher
1.9	Modified	Helen Strongman	Shivani Padmanabhan
1.10	Modified	Helen Strongman	Shivani Padmanabhan
1.11	Modified	Rebecca Ghosh	Shivani Padmanabhan
1.12	Modified	Rebecca Ghosh	Arlene Gallagher
1.13	Modified	Arlene Gallagher, Rebecca Ghosh	Shivani Padmanabhan
1.14	Modified	Susan Hodgson	Rachael Williams
1.15	Modified	Susan Hodgson	Kirsty Syder
1.16	Modified	Susan Hodgson	Achim Wolf
1.17	Modified	Susan Hodgson	Hilary Shepherd
1.17.1	Modified	Susan Hodgson	Eleanor Yelland
1.18	Modified	Susan Hodgson	Suhail Shiekh
1.19	Modified	Susan Hodgson	
1.20	Modified	Susan Hodgson	
1.30	Modified	Jenny Campbell	Justin Chan

### Summary of Changes

#### Version 1.1

- Modified linkage coverage dates for Cancer Registry data

#### Version 1.2

- Modified example for estimating end of follow up

#### Version 1.3

- Modified explanation of linkdate and inclusion/exclusion criteria for linkage eligibility file
- Modified eligibility criteria for ONS death data post linkage set 7
- Added information on the linkage coverage and linked practice file

#### Version 1.4

- Modified eligibility criteria for MINAP data post linkage set 8

#### Version 1.5

- Updated for linkage set 10
- Eligibility criteria include read codes for dissent that have been added to the dictionary

- Eligibility criteria for MINAP and Cancer registry data
- Information about the new eight step matching process
- Linkage coverage periods are now provided by the data custodian

#### Version 1.6

- Updated header and footer with new agency branding

#### Version 1.7

- Modified explanation of patients who are not included in the linkage source file
- Clarified application of HES eligibility flag to all HES data sources
- Eligibility criteria for Cancer Registry data, MINAP data and LSOA data
- The restriction to records with match rank  $\leq 5$  refers to the linked records that are provided and not to whether or not the patient is considered eligible for linkage
- The linked practice file will no longer be provided as standard. This is to ensure that the linkage source file is used to assess linkage eligibility. This only includes patients registered in practices contributing to the linkage programme prior to the linkdate (see below)
- Modified description of definition of coverage start and end dates
- Extended recommendations for identifying linked study population

#### Version 1.8

- Added explanation about retained data for records matched using steps 6-8
- Further clarification of application of HES eligibility flag to all data sources

#### Version 1.9

- Added explanation about difference between steps 6 and 7 of the matching algorithm

#### Version 1.10

- Updated for linkage set 13
- Changed name of HSCIC to NHS Digital following rebranding
- Patients with type 2 objections now removed by NHS Digital
- Additional description of eligibility flag to include PROMs

#### Version 1.11

- Updated for linkage set 14
- Additional eligibility flag added for Mental health data
- Clarified description of cr\_e eligibility flag

#### Version 1.12

- Updated for linkage set 15
- Updated text on the linkdate changes
- Removed MINAP eligibility flag and text
- Updated header and footer with new agency branding

#### Version 1.13

- Updated for linkage set 16
- Updated to include CPRD Aurum

#### Version 1.14

- Updated for linkage set 17

#### Version 1.15

- Updated for linkage set 18

#### Version 1.16

- Updated for linkage set 19

#### Version 1.17

- Updated for linkage set 20

#### Version 1.17.1

- Updated to add reference to the CPRD GOLD 'linkage\_eligibility\_new\_patids.txt' which contains the new five digit practice ID numbering convention, which also affects the patient ID number, and which was applied from the Jan 2021 build.

#### Version 1.18

- Updated for linkage set 21
- DOI added

#### Version 1.19

- Updated to edit reference to ISAC

#### Version 1.20

- Updated to reflect move to provision of latest linked data available per linked dataset via enhanced eligibility file

#### Version 1.30

- Updated to include Nov 2024 update to HES APC, HES OP, ONS Death and Small Area Level Data. This version includes a change in linkage methodology for these datasets and a significant change in the definition of eligibility.

## **What is the source file (GOLD\_enhanced\_eligibility\_[month]\_[year].txt or Aurum\_enhanced\_eligibility\_[month]\_[year].txt)?**

The source data are provided for users to select the patients that are eligible for linkage between the CPRD GOLD or CPRD Aurum datasets and the available linked data sources. All patients in the source file are from practices based in England<sup>1</sup> that have consented to take part in the linkage process and have not opted-out or dissented from the sharing of confidential patient information (such as the identifiers required for linkage) for planning and research. A separate linkage eligibility file is provided for CPRD GOLD and CPRD Aurum - these are identical in format. The linkage eligibility files (or 'source' files) contain flags to designate whether the patient is eligible for linkage to each individual data source. Some patients will not be eligible for any of the linkages, whereas others may be eligible for some/all of them. These data are provided in order to support researchers to make a decision on the denominator population for a study.

The variable linkyear reflects the year in which the cohort of patients represented in this source file was sent to the trusted third party for linkage. Patients registered for the first time in a practice after the linkyear will not be in the source file although they may be in the CPRD GOLD or CPRD Aurum builds processed after the identifiers were sent to the trusted third party.

The trusted third party attempts to link all patients with the exception of those who have opted out or dissented from providing data to CPRD for research or from disclosure of personal confidential data to NHS England whilst registered at any UK practice.

## **Linkage methods**

### **Master Person System (MPS)**

As of the November 2024 linked data update linkage of HES APC, HES OP and the ONS Death Data to CPRD primary care data is done using NHS England's Master Person Service (MPS) which replaces the eight-step deterministic algorithm used previously. The Master Person Service extracts demographic information from the Personal Demographics Service database (PDS) held by NHS England, which has approximately 80 million patient records. CPRD patients are matched to PDS patients on NHS number and date of birth. If patients cannot be matched to the PDS database, then the MPS process attempts to ascertain a match based on date of birth, gender and either a local patient identifier or postcode to records in the MPS record bucket (~6% of linked patients in CPRD GOLD and ~2% of linked patients in CPRD Aurum were matched using this secondary method). Each matched patient is assigned a CPRD specific MPS identifier (CPRD MPSid). A list of matched patients' CPRD MPSid's together with their data is provided to CPRD. Further information on how the Master Person Service works can be found in NHS England's Person ID Handbook (1).

### **Eight step linkage algorithm**

---

<sup>1</sup> CPRD can only link patients from English practices because our trusted third party, NHS Digital who conduct the linkage, has responsibility for data and information from across the health and social care system in England, not the devolved nations.

For datasets which were not updated as part of the November 2024 linked data release (HES A&E, HES DID, SGSS, CHESS and NCRAS data) the eight-step algorithm remains the linkage method. The trusted third party use an eight-step process to match patients using some or all of the following: NHS number, date of birth, sex and postcode. The table below outlines the eight steps:

Table 1: NHS England 8 step linkage algorithm

Step	NHS Number	Date of Birth	Sex	Postcode
1	Exact	Exact	Exact	Exact
2	Exact	Exact	Exact	
3	Exact	Partial	Exact	Exact
4	Exact	Partial	Exact	
5	Exact			Exact
6*		Exact	Exact	Exact
7**		Exact	Exact	Exact
8	Exact			

\* where NHS number doesn't contradict the match, DOB not 1st of January & postcode not on the communal establishment list

\*\* where NHS number doesn't contradict the match and DOB is not 1st of January

Each of these data sources includes a match\_rank variable, indicating the step at which the match between a patient in CPRD and each data source was established. For each linked data source, the documentation includes tabulations showing the relative frequency of record matching at each step. Only records matched using steps 1-5 are provided as standard. Records matched on steps 6-8 have been retained in separate files. Modified linkage eligibility files can be made available; we envisage that the retained records and modified eligibility files will primarily be of interest to methodological researchers. Please speak to a member of the CPRD Observational Research team prior to submission of your protocol to request these data.

### Definition of eligibility

The source file contains flags to designate whether the patient is eligible for linkage to each individual data source. With the November 2024 update the definition of eligibility is different depending on the linkage method used.

For data linked using the MPS system (currently HES APC, HES OP and ONS Death data) patients are flagged as eligible if they have successfully been linked to an CPRD MPSid. This means for example that patients who are flagged as eligible but do not have HES data have not been to hospital.

- **ONS Death** (ons\_death\_e = 1): Patients registered at a practice participating in the linkage scheme, successfully linked using the MPS system,
- **HES** (hes\_apc\_e = 1, hes\_op\_e = 1): Patients registered at a practice participating in the linkage scheme, successfully linked using the MPS system,

For small area data linked via patient postcode patients are now flagged as Isoa\_e= 1 only if they were successfully linked.

- **Small area level data based on LSOA** (Isoa\_e = 1): Patients registered at a practice participating in the linkage scheme, with a valid postcode, who have been successfully linked to small area level data based on English lower super output area (LSOA), including the Index of Multiple Deprivation (IMD), Carstairs Index and Townsend Score, and Rural-Urban Classification.

For data linked using the eight-step algorithm being eligible for linkage does not mean that a linkage was successfully made, just that the patient had the correct information to be sent for potential linkage.

Amongst patients eligible for linkage for HES A&E, HES DID, SGSS, CHES and NCRAS data, lack of linked data may reflect there being no linked data for that patient, or that linkage was not successful. With our next release these data will also be linked via the MPS system and the eligibility definition will be brought in line with the one above.

- **HES** (hes\_ae\_e = 1, hes\_did\_e = 1): Patients registered at a practice participating in the linkage scheme, with valid NHS number, are eligible to be linked to HES Accident and Emergency (A&E) and Diagnostic Imaging Data (DID) data.
- **COVID-19 data** (sgss\_e = 1, chess\_e = 1): Patients registered at a practice participating in the linkage scheme, with valid NHS number, are eligible to be linked to Second Generation Surveillance System COVID-19 virology test data (SGSS) and COVID-19 Hospitalisation in England Surveillance System (CHES).
- **NCRAS** (cr\_e = 1, sact\_e = 1, rtds\_e = 1): Patients registered at a practice participating in the linkage scheme, with valid NHS number, are eligible to be linked to the National Cancer Registration and Analysis Service (NCRAS) data which includes the Cancer Registration data, the Systemic Anti-Cancer Therapy Dataset (SACT), and the National Radiotherapy Dataset (RTDS)

### **What is the linkage coverage file (linkage\_coverage\_[month]\_[year].txt)?**

The linkage coverage file defines the time period each linked data source covers (start, end). The information used to define start and end dates differs for each data source.

### **Defining a linked patient cohort using CPRD GOLD and CPRD Aurum**

The following files are needed to select patients that are eligible for a study using the CPRD primary care datasets and linked data sources:

- the patient file from the relevant CPRD GOLD or CPRD Aurum build
- the practice file from the same CPRD GOLD or CPRD Aurum build
- the latest linkage eligibility file for CPRD GOLD or CPRD Aurum<sup>2</sup>
- the linkage coverage file

The CPRD GOLD practice and patient files include information that will help define the start and end of follow up for a practice (up-to-standard date (uts) and last collection date (lcd)) and the start and end of follow up for a patient (current registration date (crd), transfer out date (tod), death date (deathdate)), and the acceptability flag (accept)).

The CPRD Aurum practice and patient files include similar information that will help define the start and end of follow up for a practice (lcd but currently no uts) and the start and end of follow up for a patient (registration start date (regstartdate), registration end date (regenddate), death date (cprd\_ddate or emis\_ddate), and the acceptability flag (acceptable)).

The linkage eligibility file includes all patients who were included in the matching process and provides the indicative eligibility flags for each patient<sup>2</sup>. The linkage coverage file defines the time period each linked data source covers (start, end).

A combination of these variables should be used to:

---

<sup>2</sup> Note – in order to ensure provision of the latest available data per data source, and to honor patient opt-outs, the latest linkage eligibility per patient for each requested linked data will be applied during the processing of a linkage request. If an earlier source file was used to identify a patient cohort, this earlier eligibility information reflects indicative eligibility only. Included with the linked data provided following submission of a linkage request will be a linkage eligibility file(s) reflecting patient eligibility at the time the linkage was undertaken, and which should be used to finalise the denominator populations and associated person-time as appropriate.

- select patients who are eligible for inclusion in a study analysis
- define the time window for potential index dates and the follow-up period relative to CPRD GOLD or CPRD Aurum and linkage data collection / coverage periods
- identify denominator populations and associated person-time

For example, for a study that selects patients based on events recorded in CPRD GOLD and requires overlapping follow-up in CPRD GOLD and HES APC the following approach is recommended:

- 1) Identify study population in CPRD GOLD data using the following definition of follow-up:  
Start of follow-up = latest of crd, uts, start of HES APC coverage, start of study period  
End of follow-up = earliest of tod, deathdate, lcd, end of HES APC coverage, end of study period
- 2) Exclude patients with no follow-up time (end of follow-up < start of follow-up)
- 3) Merge with linkage eligibility file and select matched patients who are eligible for HES APC linkage (hes\_apc\_e = 1)

For the equivalent example in CPRD Aurum the following approach is recommended:

- 1) Identify study population in CPRD Aurum data using the following definition of follow-up:  
Start of follow-up = latest of regstartdate, start of HES APC coverage, start of study period  
End of follow-up = earliest of regenddate, cprd\_ddate, lcd, end of HES APC coverage, end of study period
- 2) Exclude patients with no follow-up time (end of follow-up < start of follow-up)
- 3) Merge with linkage eligibility file and select matched patients who are eligible for HES APC linkage (hes\_apc\_e = 1)

Included with the linked data provided following submission of a linkage request will be a linkage eligibility file(s) reflecting patient eligibility at the time the linkage was undertaken – these data should be used to finalise the denominator populations and associated person-time as appropriate for the study.



**Linkage Source (GOLD\_enhanced\_eligibility\_[month]\_[year].txt or Aurum\_enhanced\_eligibility\_[month]\_[year].txt)**

<i>Column Name</i>	<i>Description</i>	<i>Type</i>	<i>Format</i>
patid	The encrypted unique identifier given to a patient in CPRD GOLD or CPRD Aurum [primary key]	CHAR	20
pracid	The encrypted unique identifier given to a practice in CPRD GOLD or CPRD Aurum	INTEGER	5
linkyear	The year in which the information of the patient was sent to the trusted third party for linkage	DATE	yyyy
hes_apc_e	Flag (0,1) indicating whether patient is eligible for linkage to HES APC data	INTEGER	1
ons_death_e	Flag (0,1) indicating whether patient is eligible for linkage to ONS death registration data	INTEGER	1
lsoa_e	Flag (0,1) indicating whether patient is eligible for linkage to patient small area level data based on LSOA	INTEGER	1
sgss_e	Flag (0,1) indicating whether patient is eligible for linkage to SGSS COVID-19 data	INTEGER	1
chess_e	Flag (0,1) indicating whether patient is eligible for linkage to CHESS COVID-19 data	INTEGER	1
hes_op_e	Flag (0,1) indicating whether patient is eligible for linkage to HES OP data	INTEGER	1
hes_ae_e	Flag (0,1) indicating whether patient is eligible for linkage to HES A&E data	INTEGER	1
hes_did_e	Flag (0,1) indicating whether patient is eligible for linkage to HES DID data	INTEGER	1
cr_e	Flag (0,1) indicating whether patient is eligible for linkage to NCRAS Cancer Registration data	INTEGER	1
sact_e	Flag (0,1) indicating whether patient is eligible for linkage to NCRAS SACT data	INTEGER	1
rtds_e	Flag (0,1) indicating whether patient is eligible for linkage to NCRAS RTDS data	INTEGER	1

**Linkage Coverage (linkage\_coverage\_[month]\_[year].txt)**

<i>Column Name</i>	<i>Description</i>	<i>Type</i>	<i>Format</i>
data_source	Linkage data source	CHAR	25
start	The start date of collection of data for this source	DATE	dd/mm/yyyy
end	The end date of collection of data for this source	DATE	dd/mm/yyyy

**Reference**

- (1) Data Science Team NHS England, 'The Person\_ID Handbook', 2024. Accessed: Nov. 08, 2024. [Online]. Available: [https://digital.nhs.uk/services/personal-demographics-service/master-person-service/the-person\\_id-handbook](https://digital.nhs.uk/services/personal-demographics-service/master-person-service/the-person_id-handbook)