Medicines & Healthcare products Regulatory Agency

CPRD

Linkage Source Data Documentation (set 21)

Version 1.19

Date: 28 May 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	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	

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 <=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

What is the source file (linkage_eligibility.txt or linkage_eligibility_new_patids.txt[1])?
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 linkage_eligibility.txt/linkage_eligibility_new_patids.txt file are from practices based in England[2] 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 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.

In addition to eligibility flags, for each patient a linkdate is also provided. This outlines the date when the key information needed for linkage was sent by the practice to the trusted third party. The linkdate does not affect the availability of person-time in the CPRD primary care data or the linkage data; it is merely an indicator of when the NHS number, date of birth, sex and postcode was sent to the trusted third party for linkage. The linkdate may not be the same as the practice last collection date for the latest build. Prior to set 15, the linkdate was set to the same date for all patients within an individual practice. From set 15 it is possible for patients registered at the same practice to have different linkdates. This affects a very small number of patients, and occurs when a practice has subsequently split into more than one practice.

Patients registered for the first time in a practice after the linkdate 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 Digital whilst registered at any UK practice. Note that being eligible for linkage does not mean that a linkage was successfully able to be made. Amongst patients eligible for linkage for each source, lack of linked data may reflect there being no linked data for that patient, or that linkage was not successful. CPRD is provided a list of patients and information on the linkage matching process.

Definition of eligibility

The source file contains flags to designate whether the patient is eligible for linkage to each individual data source.

- **HES** (hes_e = 1): Patients registered at a practice participating in the linkage scheme, with valid NHS-number, are eligible to be linked to HES data (Admitted Patient Care), Outpatient, Accident and Emergency (A&E), Diagnostic Imaging Data (DID) and Patient Reported Outcomes Measures (PROMS).
- **ONS Death** (death_e = 1): Patients registered at a practice participating in the linkage scheme, with valid NHS-number, are eligible to be linked to ONS death registration data.
- **NCRAS** (cr_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

[1] From the January 2021 build, CPRD GOLD have used the new five digit practice ID numbering convention, which also affects the patient ID number. The 'linkage_eligibility.txt' file contains the old format (i.e. pre Jan 2021) patids and pracids; the 'linkage_eligibility_new_patids.txt' file contains the new format (i.e. Jan 2021 onwards) patids and pracids. Customers are advised to check which file should be used to ensure they are consistent in their use of old format or new format versions of pracid and patid across build and source file.
[2] 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.

(NCRAS) data which includes the Cancer Registration data, the Systemic Anti-Cancer Therapy Dataset (SACT), the English Cancer Patient Experience Survey (CPES), the National Radiotherapy Dataset (RTDS), the Quality of Life of Cancer Survivors in England: Pilot Survey (QOLP) and the Quality of Life of Colorectal Cancer Survivors in England: Patient Reported Outcome Measures Survey (QOLC).

- **Mental Health** (mh_e = 1): Patients registered at a practice participating in the linkage scheme, with valid NHS-number, are eligible to be linked to Mental Health data.
- **Small area level data based on LSOA** (lsoa_e = 1): Patients registered at a practice participating in the linkage scheme, with a valid postcode, are eligible to be 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 socioeconomic scores, and Rural-urban classification. A postcode is considered valid if it is in the correct format (length and characters in the correct place). There is therefore the potential for data to be missing if the postcode or LSOA could not be matched to the reference data. Further information is available in the documentation on small area data.

Please note: For the first CPRD Aurum linked dataset (set 16), the trusted third party were only provided information for patients with valid NHS numbers, therefore the proportion of eligible patients to ineligible patients in CPRD Aurum is artificially high when compared to CPRD GOLD. From set 17, patients with invalid NHS numbers were also provided, and the proportion of eligible to ineligible patients is now more closely aligned between CPRD Aurum and GOLD.

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 Digital 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 data source 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; matching on steps 6 and 7 do not require an NHS number. 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.

Eligibility in set 21

Eligibility for CPRD GOLD and CPRD Aurum for linkage set 21 are summarised in Tables 2A and 2B below.

Table 2A: Number of **CPRD GOLD** patients in the source file, and with key identifiers for eligibility for linkage in set 21.

Number of patients in source file	11,031,736
Number of patients with valid NHS number as recorded in their primary care record[1]	8,880,055
Number of patients with valid postcode as recorded in their primary care record[2]	10,524,262
Number of acceptable patients in Jan 2021 build eligible for >/=1 linkage	9,268,968

[1] Patients in the source file with a valid NHS number are eligible for linkage to HES, ONS Deaths, NCRAS data and Mental Health data

[2] Patients in the source file with a valid postcode are eligible for linkage to LSOA and accompanying small area data

Table 2B: Number of **CPRD Aurum** patients in the source file, and with key identifiers for eligibility for linkage in set 21.

Number of patients in source file	47,396,378
Number of patients with valid NHS number as recorded in their primary care record [1]	38,932,601
Number of patients with valid postcode as recorded in their primary care record [2]	45,045,090
Number of acceptable patients in Jan 2021 build eligible for >/=1 linkage	37,714,624

[1] Patients in the source file with a valid NHS number are eligible for linkage to HES, ONS Deaths, NCRAS data and Mental Health data

[2] Patients in the source file with a valid postcode are eligible for linkage to LSOA and accompanying small area data

DOI
Please cite in any publications using these data:

CPRD GOLD Source file March 2021 (set 21) - https://doi.org/10.48329/77ws-ev28

CPRD Aurum Source file March 2021 (set 21) - https://doi.org/10.48329/537a-6q30

What is the linkage coverage file (linkage_coverage.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 linkage eligibility file for CPRD GOLD or CPRD Aurum
- 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 eligibility flags for each patient. The linkage coverage file defines the time period each linked data source covers (start, end).

A combination of these variables should be used to:

- 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 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 coverage, start of study period
   End of follow-up = earliest of tod, deathdate, lcd, end of HES 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 linkage (hes_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 coverage, start of study period
   End of follow-up = earliest of regenddate, cprd_ddate, lcd, end of HES 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 linkage (hes_e=1)

Linkage Source (linkage_eligibility.txt or linkage_eligibility_new_patids.txt [3])

Column Name	*Description*	*Type*	*Format*
patid	The encrypted unique identifier given to a patient in CPRD GOLD or CPRD Aurum [primary key]	INTEGER	20
pracid	The encrypted unique identifier given to a practice in CPRD GOLD or CPRD Aurum	INTEGER	5[4]
linkdate	Patient linkage date – the date when patient information was sent to the trusted third party (TTP) for linkage	DATE	dd/mm/yyyy
hes_e	Flag (0,1) indicating whether patient is eligible for linkage to HES data	INTEGER	1
death_e	Flag (0,1) indicating whether patient is eligible for linkage to ONS death registration data	INTEGER	1
cr_e	Flag (0,1) indicating whether patient is eligible for linkage to NCRAS data (Cancer Registration, SACT, CPES, RTDS, QOLP and QOLC)	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
mh_e	Flag (0,1) indicating whether patient is eligible for linkage to the Mental Health dataset	INTEGER	1

Linkage Coverage (linkage_coverage.txt)

Column Name	*Description*	*Type*	*Format*
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

[3] From the January 2021 build, CPRD GOLD have used the new five digit practice ID numbering convention, which also affects the patient ID number. The 'linkage_eligibility.txt' file contains the old format (i.e. pre Jan 2021) patids and pracids; the 'linkage_eligibility_new_patids.txt' file contains the new format (i.e. Jan 2021 onwards) patids and pracids. Customers are advised to check which file should be used to ensure they are consistent in their use of old format or new format versions of pracid and patid across build and source file.

[4] For GOLD linkage_eligibility.txt (i.e. pre-Jan 2021), Type/format = Integer 3