

Linking primary care data from Clinical Practice Research Datalink to secondary care and other health-related patient data: update and implications

Justin Edward Chan^{1*}, Rhys Barnett¹, Susan Hodgson¹, Prinal Chohan¹, Giulia Mantovani², and Jennifer Campbell¹

Submission History

Submitted:	20/06/2025
Accepted:	24/09/2025
Published:	19/01/2026

¹ Clinical Practice Research Datalink (CPRD), MHRA, 10 South Colonnade, Canary Wharf, London E14 4PU, UK

² NHS England, 7-8 Wellington Street, Leeds, LS1 4AP

Abstract

Introduction

Being able to accurately link primary and secondary healthcare records is invaluable for public health research. The Clinical Practice Research Datalink (CPRD) collects and curates primary care electronic health records from UK GP practices. These data are linked to secondary health data by National Health Service (NHS) England. As of 2020, NHS England introduced the Master Person Service (MPS) method to link data at the person-level. The method was first applied to CPRD data in the November 2024 linked data release.

Objectives

This paper provides an overview of the MPS linkage method and its impact on linked CPRD data.

Methods

The MPS linkage method searches each set of personal identifiers against records within the Personal Demographics Service and the MPS record bucket. Successful matches are assigned a patient identifier 'Person_ID', which is used to link records between datasets. The number of successfully linked CPRD patients was compared between the MPS and the previous linkage method. The impact of the change in linkage eligibility definition was also examined.

Results

There are 7.9 million (CPRD GOLD) and 34.2 million (CPRD Aurum) patient records in the December 2024 primary care builds that are of research quality and were successfully linked to a Person_ID. Compared to the previous linkage method, the proportion of patient records who were defined as eligible to be linked to Hospital Episode Statistics Admitted Patient Care (HES APC) and had data in HES APC increased from 75.7% to 81.0% in CPRD GOLD and from 72.1% to 79.0% in CPRD Aurum.

Conclusion

The new linkage eligibility definition is superior to the previous definition, resulting in greater ability to define appropriate denominator populations and to differentiate why some patients do not have linked data. The MPS linkage method offers the potential for CPRD to investigate individuals with duplicate records and practice mergers.

Keywords

electronic health records; record linkage; deterministic linkage approach; Clinical Practice Research Datalink

*Corresponding Author:

Email Address: justin.chan2@mhra.gov.uk (Justin Edward Chan)

Introduction

There is growing recognition of the utility of real-world evidence in many aspects of health research [1]. Beyond its use in drug development and drug safety monitoring, real-world evidence has the potential to play a key role in informing national guidance and healthcare decision-making [2–4]. Routinely collected Electronic Health Data is a vital tool to provide real-world evidence because of its ability to capture patient behaviour and activity in the real-world setting and over time. Patients access an array of health services for various needs, so to maximise the utility of these electronic health records (EHR) it is necessary to link multiple EHR datasets together to give a more complete picture of a patient's healthcare journey.

The Clinical Practice Research Datalink (CPRD) is a UK Government research service jointly supported by the Medicines and Healthcare products Regulatory Agency (MHRA) and the National Institute for Health Research (NIHR) to promote public health and clinical studies using UK patient EHR. CPRD collects primary care patient EHR from UK GP practices and, for patients in England, receives linked patient-level health data representing secondary and specialist care, and mortality data, from National Health Service (NHS) England. These linked data sources complement primary care records and together form a more complete UK population health dataset [5]. Researchers across academic, regulatory, and pharmaceutical organisations worldwide have used CPRD data to conduct observational public health research [6]. In the past five years, over 88% of protocols submitted to CPRD requested a combination of patients' primary care data and their linked secondary care data.

In accordance with the current NHS research ethics committee approval, CPRD has access to pseudonymised patient data that does not include NHS number or full date of birth (DOB) [7]. Instead, NHS England (NHSE; previously known as NHS Digital), acting as a Trusted Third Party, performs data linkage on behalf of CPRD in line with existing legal support from the UK's Health Research Authority (HRA) Confidentiality Advisory Group (CAG) [8]. As of 2020, NHSE commenced work to change the method used to link data across different systems and datasets from a deterministic eight-step algorithm [9] to a deterministic approach using the NHSE Master Person Service (MPS) [10]. The change was instigated by NHSE to standardise data linkage across different datasets and improve the ability to link incomplete records to the correct person [11].

The objective of this paper is to inform researchers on this new linkage process as applied to CPRD data and its implications for public health research using CPRD primary care and linked data.

Methods

Primary care databases and data flow

General practice software, Vision® [12] and EMIS® [13], support General Practitioners (GPs) to record all symptoms, clinical measurements, laboratory test results, diagnoses, and prescription events during routine clinical care. GP practices

can elect to contribute the fully coded patient EHR to CPRD for their patients. Data are pseudonymised, with direct patient identifiers replaced with pseudonymised system patient and practice identifiers before transferring the coded medical data to CPRD. Type 1 opt-out and the National Data Opt-Out (NDOO) preferences are honoured between the software suppliers and CPRD, such that data from a dissented patient do not flow to CPRD and no longer appear in CPRD products after their opt-out date.

Due to the different data structures and clinical coding systems between Vision® and EMIS®, CPRD processes the data as two separate primary care databases, respectively named CPRD GOLD [14] and CPRD Aurum [15]. CPRD provides updated CPRD GOLD and Aurum builds on a quarterly basis, providing a snapshot of the primary care databases for public health research. Each patient is assigned a unique patient pseudonym (CPRD patid) when they join a practice but are allocated a new pseudonym if they move to another practice or if their practice is absorbed by or merges with another practice [16]. In these instances, where a patient is registered at more than one CPRD contributing GP practice, the records associated with their previous patient pseudonym are retained in the database and the new patient pseudonym contains both new and historical records. This means that the same individual can potentially be included in CPRD databases multiple times.

As CPRD are not permitted to receive the patient identifiers required to link to secondary care datasets, NHSE act as a trusted third party for data linkage. Having assumed the statutory functions of NHS Digital under the Health and Social Care Information Centre (Transfer of Functions, Abolition and Transitional Provisions) Regulations 2023 [17], NHSE are legally permitted to receive identifiable patient data. Linkage is performed periodically on updated cohorts submitted to NHSE by Vision® and EMIS® to reflect active patients in currently contributing practices in England and to include practices that have only recently started contributing data to CPRD. To facilitate the linkage process, the software suppliers submit personal identifiers (NHS number, gender, DOB and postcode) and software system identifiers to NHSE. NHSE uses the MPS to validate each set of personal identifiers and return the best matched record from the Personal Demographics Service (PDS) or the MPS record bucket. Each successfully matched patient record is assigned a Person_ID, which is used by NHSE to link to their respective secondary care data. A detailed description of the MPS method is outlined in the Linkage method section.

Secondary care and other health-related data sources and data flow

CPRD offers a range of linked secondary care data sources. The full list of data sources that are currently linked to CPRD primary care data is presented in Table 1.

Data providers for each data source submit personal identifiers (NHS number, gender, DOB and postcode) and a pseudonymised patient identifier to NHSE. NHSE cannot use NHS number alone to match records. This is because some submitted records are incomplete and lack NHS numbers while others contain invalid NHS numbers or invalid combinations of

Table 1: CPRD linked data sources

Hospital Episode Statistics (HES)

HES Accident and Emergency data (HES A&E)*#
HES Admitted Patient Care (HES APC)
HES Diagnostic Imaging Dataset data (HES DID)
HES Outpatient data (HES OP)

Office for National Statistics (ONS)

Death registration data

National Cancer Registration and Analysis Service (NCRAS)

NCRAS cancer registration tumour and treatment data
NCRAS Systemic Anti-Cancer Therapy (SACT) dataset
NCRAS National Radiotherapy dataset (RTDS)

*COVID-19 Data**

COVID-19 Hospitalisation in England Surveillance System (CHESS) data
Second Generation Surveillance System (SGSS) COVID-19 positive virology test data

Small area level data

Patient and practice postcode linked deprivation measures: Index of multiple deprivation (IMD), Townsend deprivation index, Carstairs index, Rural-Urban classification

*CPRD are not receiving updated data for HES A&E, or any of the COVID-19 datasets.

The Emergency Care Data Set (ECDS) is the new national dataset for urgent and emergency care and replaced the HES A&E dataset across England from the 2019-2020 financial year. CPRD will onboard the ECDS in a future linked data refresh. Please refer to the CPRD website for the latest updates on when the ECDS will be released.

personal identifiers. As a result, NHSE uses the MPS to match the personal identifiers to their unique NHS number to confirm their identity. A bespoke deterministic linking approach is used to validate the submitted personal identifiers and provide a single best result for each record against the PDS or against the MPS record bucket. Via this process, successfully linked records from each health dataset are assigned a Person_ID. For data sources held by NHSE, after removing opted-out patients, NHSE directly submits all pseudonymised data for matched individuals in the primary care databases to CPRD.

Linkage method

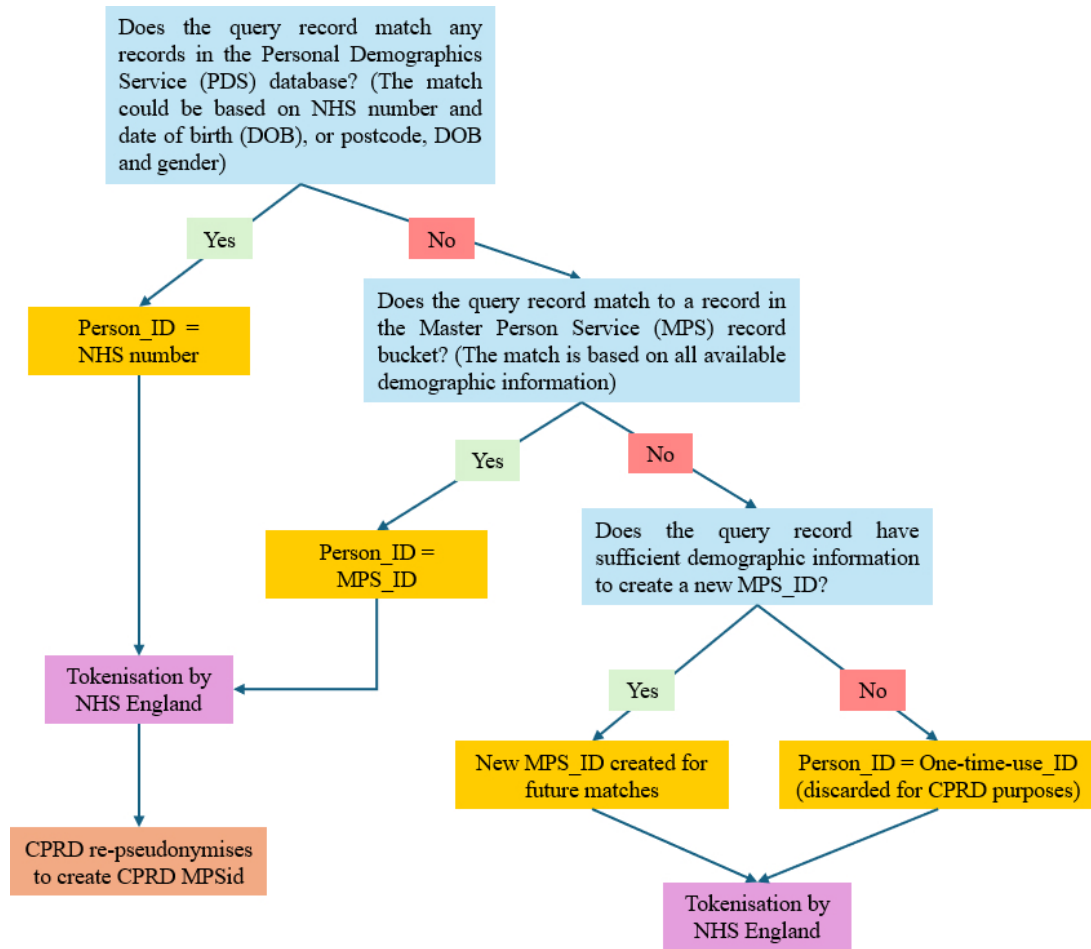
Previously, NHS England utilised an eight-stage deterministic framework to link records belonging to the same individual from different data sources [9]. The framework for this algorithm was based on eight sets of rules, each relying on a different combination of unique identifiers (e.g., NHS number and gender) and partial identifiers (e.g., DOB and postcode). The first matching step had the strictest criteria, requiring records to match on the exact NHS number, gender, DOB and postcode. The steps became progressively less restrictive. At the last matching step, records only needed to have a matching NHS number. A comprehensive description of the linkage method is provided by Padmanabhan et al. [9].

This eight-step deterministic framework was replaced by the MPS as NHSE's current method of linking records to the correct person. It operates on a different set of algorithms and searches the submitted personal identifiers against records within the PDS database and the MPS record bucket (see Figure 1). Each set of personal identifiers submitted by the data provider is referred to as a query record and a successfully matched record is assigned a Person_ID. In the first instance,

the query record is compared to the 80 million records in the PDS, the master database of NHS patients [10]. If the query record has a matching NHS number and DOB, the NHS number is used to populate the Person_ID field. If the query record contains an invalid NHS number or the DOB does not fully match with the corresponding record in the PDS, the MPS process attempts to ascertain a match based on DOB, postcode and gender. A successful match returns a Person_ID based on the NHS number of the best matched record in the PDS database. If the query record cannot be matched via the PDS database, it is compared to the records in the MPS record bucket. The MPS record bucket contains records that are unmatched in the PDS database and are allocated unique MPS identifiers (MPS_ID). The local patient identifier and other available demographic information from the query record are used to find a match from the MPS record bucket. Upon a successful match, the Person_ID field takes on the MPS_ID.

A query record that cannot be matched to the PDS database or the MPS record bucket but contains sufficient demographic information is assigned a new MPS_ID and enriches the MPS record bucket for future possible matches. If, however, a record is completely unmatchable and lacks the demographic information to generate a new MPS_ID, it is assigned a random one-time-use identifier and it is not included in the CPRD linker file. The CPRD linker file contains records with a Person_ID derived either from their NHS number from the PDS or their MPS_ID from the MPS record bucket. The Person_ID is further tokenised by NHS England, whereby the token domain is specific to CPRD, before the linker file is delivered to CPRD. CPRD re-pseudonymises the tokenised identifier to create the CPRD MPSid. The CPRD MPSid ensures CPRD users do not receive any NHS numbers and prevents further linkage to datasets prepared by different data providers.

Figure 1: Flow diagram describing MPS linkage process



Secondary care data linkage eligibility files

NHSE provides CPRD with metadata regarding the linkage process. Each record from the incoming cohort file provided by NHSE has a flag to indicate whether that patient has a valid NHS number, DOB and postcode. CPRD processes the files from NHSE and prepares two linkage eligibility files, one for CPRD GOLD and the other for CPRD Aurum. The linkage eligibility files allow CPRD users to select patients that are eligible for linkage to the various linked data sources, as this is a vital step in ascertaining the relevant denominator population for analyses that use linked data. Previously, via the eight-step linkage process, patients were considered eligible for linkage if they had the required identifiers for the linkage (i.e., NHS number in the correct format, or for small area data linkage, a valid postcode). The current linkage method changes the definition of eligibility. Patients are only considered linkage eligible by CPRD if they are successfully linked to a CPRD MPSid. For small area data, as previously, linkage eligibility requires a valid postcode.

Releasing linked data to CPRD users

Following each data linkage by NHSE, CPRD releases updated linked data to its users. Each refresh occurs periodically, with the most recent release in November 2024. It is the first time the MPS method was used to facilitate the linkage

between records and CPRD users can request the latest available linked data for patients who are successfully matched. CPRD users would typically restrict their study population to 'research acceptable' patients¹ in a chosen primary care build who are linkage eligible for the required linked data sources. To understand the impact of the MPS linkage approach, compared to the previous eight-step linkage process, the CPRD GOLD and CPRD Aurum December 2024 builds (referred to below as CPRD GOLD and CPRD Aurum) were used in the subsequent analyses.

Results

The MPS approach has led to changes in the number of patients in CPRD GOLD and Aurum linked to the secondary care data provided by NHSE. This can be observed by using the CPRD GOLD and Aurum linkage eligibility files [18, 19], which record the eligibility of each CPRD patient for linkage to the various secondary care datasets provided to CPRD by NHSE. In the most recent (November 2024) linked data release, there were 10,914,470 CPRD patient pseudonyms (CPRD patids) in the CPRD GOLD linkage eligibility file; of

¹CPRD reviews each patient's primary care record and patients with non-continuous follow-up or with poor data recording are labelled as 'research unacceptable' because the validity of the data is in question. The checks are not conducted for linked data. Please see the CPRD documentation for the list of conditions that are being checked [24, 25].

Table 2: Number of patient pseudonyms (CPRD patids) eligible for linkage in the CPRD GOLD and Aurum

	January 2022 linked data release (utilising 8-step algorithm)		November 2024 linked data release (utilising MPS linkage)	
	CPRD GOLD	CPRD Aurum	CPRD GOLD	CPRD Aurum
Number of patids in linkage eligibility file	11,067,015	48,355,828	10,914,470	43,082,693
Number of patids in linkage eligibility file eligible for HES APC linkage	8,913,996 (80.5%)	39,665,549 (82.0%)	8,690,324 (79.6%)	38,294,389 (88.9%)
From the linkage eligibility file, the number of patids that are research acceptable*	9,526,577	35,273,276	9,360,628	37,136,660
Number of patids that are research acceptable and are eligible for HES APC linkage*	8,425,296 (88.4%)	34,807,980 (98.7%)	7,939,326 (84.8%)	34,202,038 (92.1%)

*Patids with poor data recording in their primary care records are considered research unacceptable, for example, if they do not have a year of birth or are not permanently registered to a practice. The checks are not carried out for linked data.

Table 3: Comparison of HES APC eligibility flag between the November 2024 and the January 2022 linkage eligibility files (restricted to research acceptable patients in CPRD GOLD and Aurum)

CPRD GOLD (n=9,336,150)		
	HES APC eligibility flag in the January 2022 linkage eligibility file	
HES APC eligibility flag in the November 2024 linkage eligibility file	Ineligible (0)	Eligible (1)
Ineligible (0)	850,542 (9.1%)	558,730 (6.0%)
Eligible (1)	250,738 (2.7%)	7,676,140 (82.2%)
CPRD Aurum (n=35,273,275)		
	HES APC eligibility flag in the January 2022 linkage eligibility file	
HES APC eligibility flag in the November 2024 linkage eligibility file	Ineligible (0)	Eligible (1)
Ineligible (0)	346,708 (1.0%)	2,480,531 (7.0%)
Eligible (1)	118,587 (0.3%)	32,327,449 (91.6%)

these, 8,690,324 (79.6%) were eligible for HES APC linkage. For CPRD Aurum, there were 43,082,693 CPRD patids in the latest linkage eligibility file; of which 38,294,389 (88.9%) were considered eligible for HES APC linkage.

The proportion of CPRD research acceptable patids that were eligible for HES APC linkage was examined in CPRD GOLD and Aurum. Compared to the linkage eligibility files, the proportion increased from 79.6% to 84.8% in CPRD GOLD and from 88.9% to 92.1% in CPRD Aurum. Although based on similar patient cohorts supplied by Vision® and EMIS® (research acceptable patients), when directly comparing the proportion of CPRD research acceptable patids that were eligible for HES APC linkage between the previous linked data release (January 2022) which used the eight-step deterministic approach and the November 2024 linked data release (linked via the MPS), the most recent linked data release had a lower proportion of linkage eligible patids in both CPRD GOLD and Aurum (see Table 2).

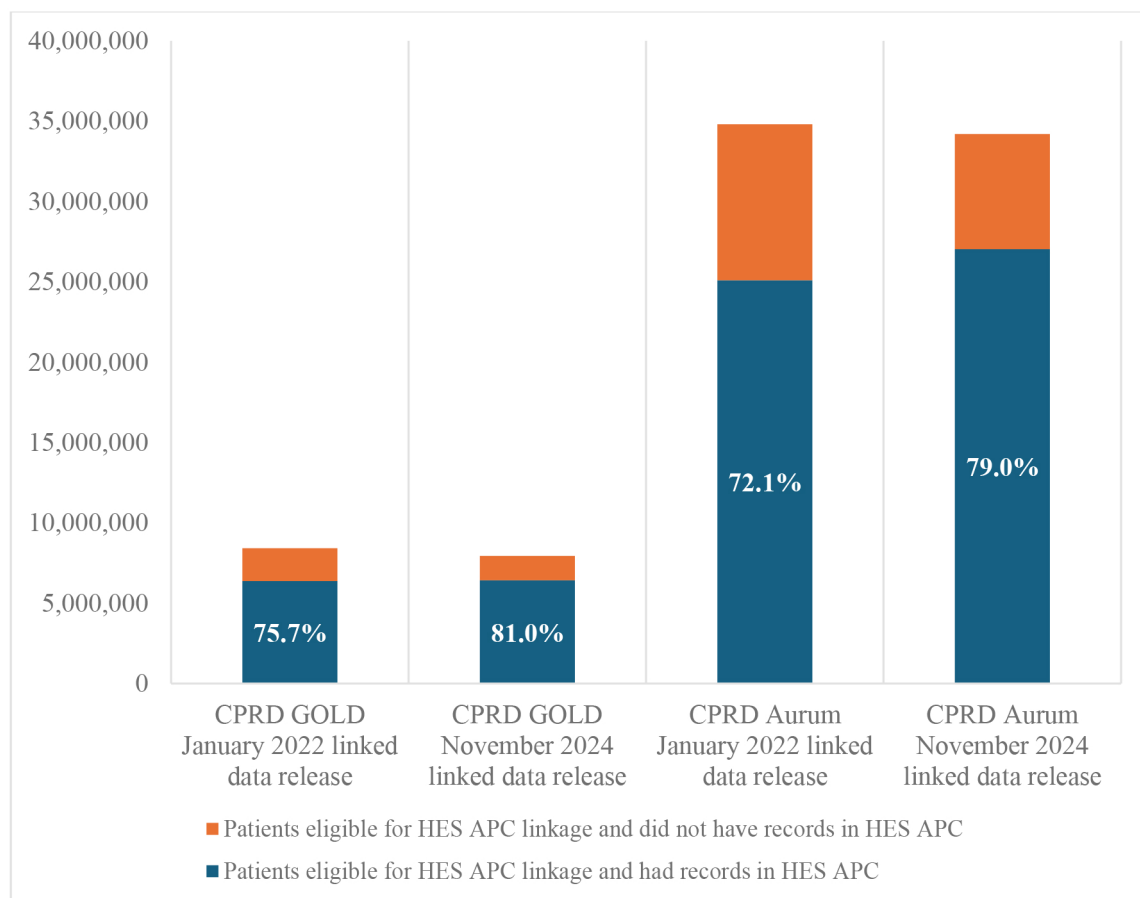
Change in definition of eligibility

The change in the definition of linkage eligibility has also contributed to a change in the proportion of eligible patients

between the most recent (November 2024) and previous (January 2022) linked data releases. Table 3 demonstrates the impact of the new eligibility definition by focusing on research acceptable patids in CPRD GOLD and Aurum that were present in both the most recent and previous linked data releases and whether their eligibility flag had changed for HES APC.

Among the CPRD GOLD patids that were present in both linked data releases, there was a 3.3% decrease in the number of patids that were linkage eligible for HES APC in the most recent linked data release compared to the previous release. However, the majority (91.3%) retained the same eligibility status for HES APC. In the previous linked data release, 6,378,980 out of 8,425,296 (75.7%) patids that were considered eligible for linkage to HES APC also had records in HES APC. The proportion increased in the most recent linked data release, where 6,431,200 out of 7,939,326 (81.0%) patids that were eligible for HES APC linkage also had data in HES APC (see Figure 2). Upon closer inspection of the 558,730 patids that were considered linkage eligible in the previous linked data release but were ineligible in the most recent linked data release, only 65,156 (11.7%) patids had HES APC records in the previous release.

Figure 2: The number/proportion of CPRD patids that were eligible for HES APC linkage and had records in HES APC (restricted to research acceptable patients in CPRD GOLD and Aurum)



The same pattern was observed in CPRD Aurum. Among the CPRD Aurum patids that were present in the two linkage eligibility files, there were approximately 2.3 million (6.7%) patids that were no longer linkage eligible for HES APC in the most recent linked data release. The eligibility status for 92.6% patids did not change. In the previous linked data release, 25,102,136 out of 34,807,980 (72.1%) patids were considered linkage eligible for HES APC and had records in HES APC. Like CPRD GOLD, the proportion also increased in the most recent linked data release, where 27,036,545 out of 34,202,038 (79.0%) patids that were eligible for HES APC linkage had records in HES APC (see Figure 2). When examining the 2,480,531 patids that were considered eligible for HES APC linkage in the previous linked data release but were ineligible in the most recent linked data release, only 168,316 (6.8%) had HES APC records in the previous release.

The relationship between CPRD MPSid and CPRD patient pseudonym

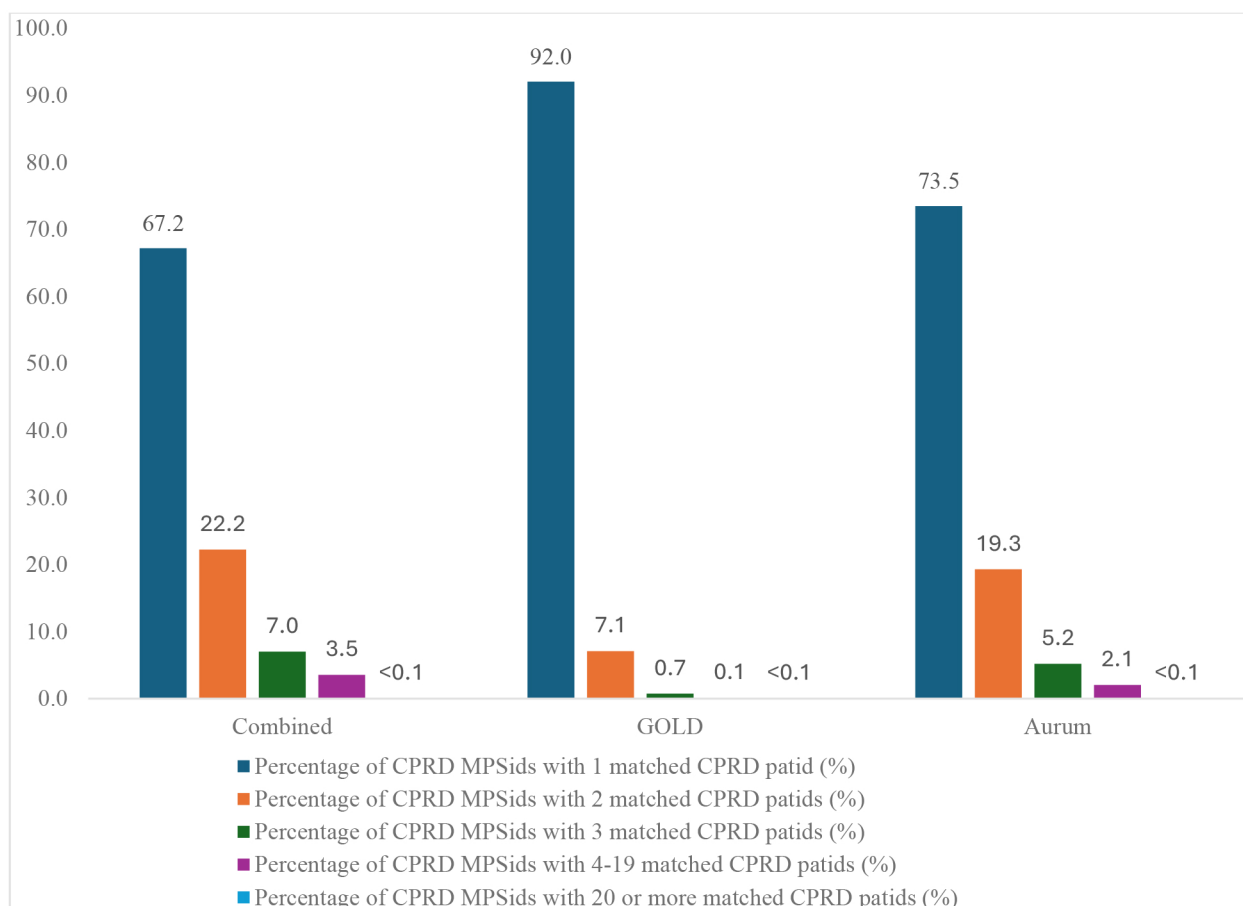
An individual can be assigned more than one CPRD patid if they move between contributing practices or if their original practice is absorbed by or merges with another practice. Via the previous eight-step deterministic linkage process, linked patients who had HES records were assigned an individual-specific HES pseudonym (HES_id). For individuals who had not been hospitalised (or whose hospital data could not

be linked), the absence of this HES pseudonym made it challenging to determine whether CPRD patids with similar medical records belonged to the same individual.

The MPS linkage approach can potentially be used internally by CPRD to discern data that is contributed by the same person across multiple practices because CPRD patids from the same person share the same CPRD MPSid. Patients linked via the MPS are assigned an CPRD MPSid irrespective of whether they appear in hospital records. Among patients successfully matched to a CPRD MPSid, 21,205,347 (67.2%) CPRD MPSids in CPRD GOLD and Aurum had a 1:1 match with a CPRD patid. Of those CPRD MPSids that had a 1:1 match, 3,368,742 had a match with a CPRD patid in CPRD GOLD and 17,836,605 with a CPRD patid in CPRD Aurum. As illustrated in Figure 3, the majority of CPRD MPSids had between one and three associated CPRD patids (96.4%). There was a small proportion of CPRD MPSids that had more than 20 matched CPRD patids (<0.1%).

Discussion

The MPS linkage method is part of NHSE's strategy to improve their data processing services [20]. It standardises the way patient records from different datasets are linked, thereby improving the likelihood of attributing data to the correct individual [10]. The ability to accurately link patient records across different datasets is necessary to provide researchers

Figure 3: Histogram illustrating the number of CPRD MPSids matched to CPRD patids²

with reliable and useful data to address important public health research questions. The MPS linkage method uses the national master database of all NHS patients in England, the PDS database, which has complete coverage of patients who have interacted with an NHS care setting. Each patient record in the PDS database also contains an array of useful non-clinical information, such as historic and current postcodes, which maximises the chance of a successful match. Furthermore, the live version of the PDS database is queried during the MPS linkage process. This ensures that all patient records matched in the PDS database contain up to date information (e.g., their latest postcode). It is important to note that this does not apply to query records matched in the MPS record bucket; records in the bucket are not updated [10].

The use of linkage to a Person_ID to define linkage eligibility is superior to the former approach which indicated the availability of the identifiers required for linkage, rather than the result of the linkage itself. Previously, linkage eligible individuals without any linked data could be attributed to a person never interacting with an NHS service (i.e., genuine absence of linked data) or due to that patient having invalid identifiers in their GP record (e.g., incorrect NHS number or DOB). This inability to disentangle these different scenarios affected the ability to define the appropriate denominator populations, potentially impacting prevalence and incidence rates. The new definition of linkage eligibility provides CPRD

users with more clarity in instances where individuals do not have any linked data. For example, if a patient is linkage eligible for HES APC (i.e., successfully matched under the MPS linkage method) but does not have any HES APC records, this is because they have never been admitted to a hospital. Conversely, if a patient is linkage ineligible for HES APC, they do not have HES APC records because they lacked valid identifiers to be matched under the MPS linkage method.

The change in the linkage eligibility definition has impacted the numbers of patients in the linkage eligibility files between the November 2024 and January 2022 linked data releases. There was a noticeable number of CPRD research acceptable patids from CPRD GOLD and Aurum that were previously considered eligible for HES APC linkage but did not have a CPRD MPSid (Table 2). Among the CPRD patids that were linkage eligible in the previous release, but not in the most recent release (CPRD GOLD n=558,730; CPRD Aurum n=2,480,531; see Table 3), 92.3% (n=2,805,789) did not have any HES APC records in the January 2022 linked data release, either because they had not been admitted to hospital, or their identifiers such as NHS numbers were incorrect. This suggests the previous linkage eligibility definition may have overestimated the number of patients who were eligible for linkage. It is plausible to speculate that a number of these patients had incorrect identifiers to be successfully linked to records in HES APC. A small portion of CPRD patids that were considered linkage ineligible for HES APC in the previous linked data release now have a CPRD MPSid. This may be

²CPRD recommends researchers to remove patients with a CPRD MPSid that is matched to 20 or more CPRD patids from their analyses.

partially explained by the MPS linkage approach having access to more complete and robust patient records from the PDS database and the MPS record bucket, therefore improving the ability to ascertain a match. Overall, the new linkage eligibility definition saw an increase in the proportion of patients considered linkage eligible and have records in HES APC (see Figure 2). This provides support that using the CPRD MPSid as an eligibility criterion is more rigorous at removing patients with incorrect identifiers from the denominator population.

The CPRD MPSid may potentially be used to rationalise follow-up time for patients who have moved between practices. While an individual can be assigned multiple CPRD patids, they can only assume one CPRD MPSid. CPRD may be able to use CPRD MPSids, alongside registration dates, to remove duplicate records and curate a single record for each patient with longer follow-up time. At present, researchers tend to restrict follow-up time to currently registered periods to avoid overestimation of incidence rates [21]. Having access to a patient's record for a longer period time may help identify past events that are being recorded during the first few months of registration to a new practice and avoid contemporaneous events from being censored.

Similarly, CPRD may use the CPRD MPSid to gain further insight into practice mergers by exploring the overlap of CPRD MPSids across multiple practices. This may enable identification of practices that have merged, allowing patient records from practices that are duplicated to be correctly removed from the CPRD GOLD and Aurum databases.

There remain some challenges when using the MPS linkage method. While it represents an improvement compared to the previous linkage approach adopted by NHS England, it still has limitations. Specifically, MPS is highly reliant on DOB. A missing DOB causes the linkage to automatically fail. A record with a wrong or partially wrong DOB could be matched only in a restricted number of cases. The adoption of the CPRD MPSid as the new definition of linkage eligibility has therefore potentially altered the composition of the cohort eligible for data linkage. Certain groups, such as patients with naming variations or patchy records may experience more linkage problems and it is unclear if the MPS method of linkage excluded these individuals more or less frequently than the previous method. Future work to examine demographic differences between those who were retained or lost by the MPS method of linkage would be useful to evaluate the representativeness of the data. Furthermore, since the MPS method of linkage queries the live version of the PDS database and can replace outdated and invalid identifiers with the latest and correct versions, it would be interesting to explore which identifiers (e.g., DOB, NHS number, gender and postcode) contributed most to the extra matches.

The most recent linked data release also reveals a small number of CPRD MPSids having many matched CPRD patids (see Figure 3). While this may reflect the MPS being able to capture individuals with higher frequency of movement and changes in GP practice, this may also arise from the MPS process when it is unable to distinguish individuals with limited but similar identifier values (Case study 15 in the Person_ID handbook) (10). These records may not be reliable and CPRD users with access to linked datasets should consider using the CPRD MPSid to exclude records associated with 20 or more CPRD patids from their analysis. CPRD may utilise available

demographic information of these CPRD patids in the future to evaluate their validity and discern if they are from the same patient.

Conclusion

The MPS linkage approach was applied for the first time in the context of CPRD data to the November 2024 linked data release, which included HES APC, HES OP, ONS death registration data and small area data. Compared to the previous approach, the new process has improved the ability to link data from various datasets to the correct individual. The change in linkage eligibility definition allows CPRD users to better prepare their study and denominator populations by removing patients who are ineligible to be linked to their desired dataset. Future linked data releases will utilise the MPS linkage approach. CPRD welcomes validation studies and collaborative projects to ensure the robustness of CPRD linked data in addressing public health research.

Acknowledgement

We thank John Wigglesworth for his assistance in understanding the impact of the MPS linkage method on CPRD data. We also acknowledge the support of Linda Wijlaars, who provided helpful feedback on an earlier version of the manuscript.

Statements of conflicts of interest

Justin Chan, Rhys Barnett, Susan Hodgson and Jennifer Campbell declare that this work was conducted during their current employment at the Clinical Practice Research Datalink (CPRD). Prinal Chohan declares this work was conducted during her former employment at CPRD. Giulia Mantovani is the Head of Data Linkage Hub, NHS England and contributed to this work during their current employment at NHS England. The authors have read the ICJME Form for Disclosure of Potential Conflicts of Interest and have no relevant financial or non-financial interest to disclose.

Ethics statement

This study uses anonymised electronic health records only. Approval was obtained from the East Midlands - Derby Research Ethics Committee (REC reference 21/EM/0265).

Data availability statement

This study is based in part on data from the Clinical Practice Research Datalink obtained under licence from the UK Medicines and Healthcare products Regulatory Agency (MHRA). The data is provided by patients and collected by the NHS as part of their care and support. Hospital Episode Statistics (HES) data are copyright © 2025 and re-used with the permission of The Health & Social Care Information Centre. All rights reserved. The interpretation and conclusions contained in this study are those of the authors alone.

The data that support the findings of this study are available from CPRD, but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Requests to access CPRD data are reviewed via the CPRD Research Data Governance process to ensure that the proposed research is of benefit to patients and public health. More information is available on the CPRD website: <https://www.cprd.com/safeguarding-patient-data>. This study utilised data from the December 2024 build of CPRD GOLD [18] and CPRD Aurum [19] with linked data from HES APC from the November 2024 linked data release [22, 23]. Upon reasonable application to the CPRD RDG, researchers may use this information to assemble the data used in this study. For further information, please contact the study authors in the first instance.

Funding statement

CPRD is jointly sponsored by the MHRA and the National Institute for Health Research (NIHR). As a not-for-profit UK government body, CPRD seeks to recoup the cost of delivering its research services to academic, industry, and government through fees. The interpretation and conclusions contained in this study are those of the authors alone. For further information, please contact the study authors in the first instance.

Authors' contributions

Justin Chan and Jennifer Campbell contributed to the study conception and design. Data preparation was carried out by Rhys Barnett and Prinal Chohan. The analysis was performed by Justin Chan and all authors facilitated the interpretation of the results. The first draft of the manuscript was written by Justin Chan and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

References

- Dang A. Real-World Evidence: A Primer. *Pharmaceut Med.* 2023 Jan 5;37(1):25–36. <https://doi.org/10.1007/s40290-022-00456-6>
- National Institute for Health and Care Excellence. NICE real-world evidence framework [Internet]. 2022 [cited 2024 Nov 8]. Available from: <https://www.nice.org.uk/corporate/ecd9/chapter/overview>
- Zisis K, Pavi E, Geitona M, Athanasakis K. Real-world data: a comprehensive literature review on the barriers, challenges, and opportunities associated with their inclusion in the health technology assessment process. *Journal of Pharmacy & Pharmaceutical Sciences.* 2024 Feb 28;27. <https://doi.org/10.3389/jpps.2024.12302>
- Schad F, Thronicke A. Real-World Evidence—Current Developments and Perspectives. *Int J Environ Res Public Health.* 2022 Aug 16;19(16):10159. <https://doi.org/10.3390/ijerph191610159>
- Clinical Practice Research Datalink. Clinical Practice Research Datalink [Internet]. 2024 [cited 2024 Nov 25]. Available from: <https://www.cprd.com/>.
- Clinical Practice Research Datalink. Bibliography [Internet]. 2024 [cited 2024 Nov 25]. Available from: <https://www.cprd.com/bibliography>.
- Health Research Authority. Clinical Practice Research Datalink (CPRD) Research Database - Health Research Authority [Internet]. 2022. [cited 2024 Nov 25]. Available from: <https://www.hra.nhs.uk/planning-and-improving-research/application-summaries/research-summaries/clinical-practice-research-datalink-cprd-research-database/>.
- NHS England. Data [Internet]. [cited 2024 Nov 25]. Available from: <https://digital.nhs.uk/data>.
- Padmanabhan S, Carty L, Cameron E, Ghosh RE, Williams R, Strongman H. Approach to record linkage of primary care data from Clinical Practice Research Datalink to other health-related patient data: overview and implications. *Eur J Epidemiol.* 2019 Jan 15;34(1):91–9. <https://doi.org/10.1007/s10654-018-0442-4>
- Data Science Team NHS England. The Person_ID Handbook (Version 2.0.0). 2024 [cited 2024 Nov 8]. Available from: https://digital.nhs.uk/services/personal-demographics-service/master-person-service/the-person_id-handbook.
- National Health Service. Master Person Service (MPS) [Internet]. 2024. [cited 2025 May 22]. Available from: <https://digital.nhs.uk/services/personal-demographics-service/master-person-service>.
- Cegedim Healthcare Solutions. Healthcare IT Solutions for Pharmacy and Primary Care [Internet]. [cited 2024 Nov 8]. Available from: <https://www.cegedim-healthcare.co.uk/>.
- Egton Medical Information Systems. Better are through technology innovation [Internet]. [cited 2024 Nov 8]. Available from: <https://www.emishealth.com/>.
- Herrett E, Gallagher AM, Bhaskaran K, Forbes H, Mathur R, van Staa T, et al. Data Resource Profile: Clinical Practice Research Datalink (CPRD). *Int J Epidemiol.* 2015 Jun;44(3):827–36. <https://doi.org/10.1093/ije/dyv098>
- Wolf A, Dedman D, Campbell J, Booth H, Lunn D, Chapman J, et al. Data resource profile: Clinical Practice Research Datalink (CPRD) Aurum. *Int J Epidemiol.* 2019 Dec 1;48(6):1740–1740g. <https://doi.org/10.1093/ije/dyz034>
- Clinical Practice Research Datalink. CPRD Aurum Frequently asked questions (FAQs) (Version 2.6). 2024 Sep [cited 2025 May 22]. Available from: <https://www.cprd.com/sites/default/files/2024-09/CPRD%20Aurum%20FAQs%20v2.6.pdf>.

17. Health and Social Care Information Centre (Transfer of Functions, Abolition and Transitional Provisions) Regulations 2023 [Internet]. Available from: <https://www.legislation.gov.uk/uksi/2023/98/contents/made>.
18. Clinical Practice Research Datalink. CPRD GOLD Source file November 2024 (Version 2024.11.001) [Data set]. 2024 [cited 2025 May 22]. <https://doi.org/10.48329/nxvj-3672>
19. Clinical Practice Research Datalink. CPRD Aurum Source file November 2024 (Version 2024.11.001) [Data set]. 2024 [cited 2025 May 22]. <https://doi.org/10.48329/8exg-tk98>
20. NHS England. Improving our Data Processing Services (DPS) [Internet]. 2024 [cited 2024 Dec 15]. Available from: <https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/improving-our-data-processing-services>.
21. Lewis JD, Bilker WB, Weinstein RB, Strom BL. The relationship between time since registration and measured incidence rates in the General Practice Research Database. *Pharmacoepidemiol Drug Saf*. 2005 Jul;14(7):443–51. <https://doi.org/10.1002/pds.1115>
22. Clinical Practice Research Datalink. CPRD GOLD HES APC November 2024 (Version 2024.11.001) [Data set]. 2024. <https://doi.org/10.48329/fbr8-7x07>
23. Clinical Practice Research Datalink. CPRD Aurum HES APC November 2024 (Version 2024.11.001) [Data set]. 2024. <https://doi.org/10.48329/w272-7t19>
24. Clinical Practice Research Datalink. CPRD GOLD Glossary of terms v2 [Internet]. 2019 [cited 2025 Feb 20]. Available from: <https://www.cprd.com/sites/default/files/2023-02/CPRD%20GOLD%20Glossary%20Terms%20v2.pdf>.
25. Clinical Practice Research Datalink. CPRD Aurum Glossary of terms v2 [Internet]. 2019 [cited 2025 Feb 20]. Available from: <https://www.cprd.com/sites/default/files/2023-02/CPRD%20Aurum%20Glossary%20Terms%20v2.pdf>.

Abbreviations

CAG:	Confidentiality Advisory Group
CPRD:	Clinical Practice Research Datalink
DOB:	Date of birth
EHR:	Electronic health records
GP:	General Practice
HES:	Hospital Episode Statistics
HRA:	Health Research Authority
MHRA:	Medicines and Healthcare products Regulatory Agency
MPS:	Master Person Service
NDOO:	National Data Opt-Out
NHS:	National Health Service
NIHR:	National Institute for Health Research
PDS:	Personal Demographics Service

