



UK data driving real-world evidence

CPRD Aurum Frequently asked questions (FAQs)

Version 2.6

Date: 02 September 2024

Documentation Control Sheet

During the course of the project it may be necessary to issue amendments or clarifications to parts of this document. This form must be updated whenever changes are made and should be filed inside the front cover of the new or amended document.

Version	Summary of Change	Prepared/ reviewed by	Date
1.0	Initial draft	Helen Booth & Dan Dedman	06/12/2017
1.1	Updated release figures	Helen Booth	11/01/2018
1.2	Updated release figures	Helen Booth	01/02/2018
1.3	Updated release figures & linkage information	Helen Booth	01/03/2018
1.4	Updated release figures	Helen Booth	04/04/2018
1.5	Updated release figures & linkage information	Helen Booth	02/05/2018
1.6	Updated release figures & linkage information	Helen Booth	02/07/2018
2.0	Updated for CPRD Aurum tools	Achim Wolf	10/04/2019
2.1	Full review of content	Hilary Shepherd & Arlene Gallagher	28/01/2021
2.2	Updated information on the UTS date	Helen Booth	27/04/2021
2.3	Added information about practice duplication	Dan Dedman & Mike Lonergan	10/08/2022
2.4	Referred users to the CPRD Aurum specification for information on practice mergers	Hilary Shepherd	21/12/2023
2.5	Updated section on RDG process & numunitid	Hilary Shepherd	10/05/2024
2.6	Updates sections on National Data Opt Out, comparator values and unique identifiers.	Diane Hatzioanou	15/07/2024

Contents

Contents	3
Introduction	4
What is CPRD Aurum?	4
How does CPRD Aurum differ from the CPRD GOLD database?	4
What is the population coverage of CPRD Aurum?	4
What does the CPRD Aurum database contain?	4
How can I access CPRD Aurum?	5
What is the cost for accessing CPRD Aurum data?	5
How will I know if the CPRD Aurum data are suitable for my research needs?	5
What are the differences between the CPRD Aurum and CPRD GOLD databases?	6
Points for consideration	8
Duplication of data between practices in the CPRD Aurum database	8
Duplication between the CPRD GOLD and CPRD Aurum databases	8
Medical and Product dictionaries	9
Recording of coded clinical information	9
Identifying clinical measurements	9
How are referrals recorded?	9
What is the Problem table?	9
Unique identifiers	10
Linked data	11
Guidance on applying to RDG	11
Guidance on generating code lists	11
Derived variables	12
Example research questions	13
Available results of tests, investigations, and other clinical measurements	13
Additional documentation	15
General	15
Code list generation	15

Introduction

What is CPRD Aurum?

Clinical Practice Research Datalink (CPRD) Aurum is a database of de-identified coded primary care records for use in public health research, capturing diagnoses, symptoms, prescriptions, referrals and tests. These data are contributed by general practices that use EMIS clinical systems, and the database has been named CPRD Aurum after the Latin word for 'gold'. CPRD Aurum has been available for research since 2018.

How does CPRD Aurum differ from the CPRD GOLD database?

CPRD Aurum contains data contributed by practices using EMIS clinical systems, whilst CPRD GOLD holds data from a different GP software provider (InPS Vision). Due to differences in the structure and coding of the data between the two systems the research databases have been released as separate data offerings and there are currently no plans to integrate the databases.

What is the population coverage of CPRD Aurum?

Monthly release notes are produced for the CPRD Aurum database and distributed to licence holders. They are also available through the website: <https://www.cprd.com/data-highlights>.

Practices contributing data to the CPRD Aurum database are predominantly in England and Northern Ireland; CPRD is exploring options for the inclusion of more data from across the devolved nations.

What does the CPRD Aurum database contain?

An overview of the CPRD Aurum database was published in 2019 [doi: <https://doi.org/10.1093/ije/dyz034>]. When a practice agrees to contribute data to CPRD, CPRD receives a full collection of the coded part of their electronic health records; this includes data on deceased patients and those who have left the practice. CPRD will receive all historical data, including information that was migrated from any previously used software systems. Consequently, there is some overlap between the CPRD GOLD and CPRD Aurum databases where practices have contributed via both software systems over time. Further information can be found in the '[Duplication between the CPRD GOLD and CPRD Aurum databases](#)' section.

How can I access CPRD Aurum?

Access to data from CPRD is subject to a full licence agreement containing detailed terms and conditions of use. Anonymised patient datasets can be extracted for researchers against specific study specifications, following protocol approval from the RDG (Research Data Governance) applications process. The process for CPRD Aurum data is the same as for CPRD GOLD, but applicants who have not previously used CPRD Aurum data are advised to inspect the data specifications carefully before submitting a protocol to RDG. Applicants can elect to discuss their proposals with a CPRD researcher before submitting an application, to ensure an understanding of the data structure and any implications for study design. This can be arranged via enquiries@cprd.com. Further information can be found in the '[Guidance on applying to RDG](#)' section of this document and at <https://www.cprd.com/research-applications>.

The annual multi-study licence covers both CPRD GOLD and CPRD Aurum – it is not possible to licence for these data sources separately. The CPRD online tools (Code Browser, Define and Extract) have been updated to enable access to both CPRD GOLD and CPRD Aurum. Please note that the Refine tool is only available for CPRD GOLD.

What is the cost for accessing CPRD Aurum data?

CPRD Aurum access is available through an annual multi-study licence or as a study-specific dataset, with RDG approval. Please contact enquiries@cprd.com for a quote for your specific study.

How will I know if the CPRD Aurum data are suitable for my research needs?

In the first instance, please refer to relevant publications based on CPRD Aurum data in your area of research to estimate the numbers expected. A searchable list of publications, which is updated monthly, can be found at <https://www.cprd.com/bibliography>.

Multi-study licence holders can access the Aurum database for feasibility purposes using the online tools. If your institution does not have access to the online tools, there is insufficient data in the literature and this is your first time using the Aurum data, CPRD can provide a simple feasibility count.

- Simple feasibility requests are limited to counts of patients or events recorded in a specified period.
- Simple counts should include no more than three medical and/or prescribing definitions combined, in a single request.

- Counts may be restricted to one or more of: study period, patient's age, gender and period of follow-up in CPRD Aurum.
- Counts may be stratified by calendar year, gender or age-band only.
- Users are expected to provide the relevant medical codes (Read, SNOMED, ICD-10 or OPCS codes) or therapy codes to identify events of interest in the respective data sources (code browser facilities will be provided to users).
- No denominators will be provided as part of the simple feasibility count service i.e. CPRD can provide the numerator (prevalent or incident counts), but not prevalence or incidence of an exposure or disease.
- Examples of simple feasibility counts based on 1-3 criteria are outlined in the table below.

Examples of simple feasibility counts include:

Example of counts based on <u>one</u> criterion:
1) The total number of patients in CPRD GOLD or CPRD Aurum with a first ever <u>prescription</u> for metformin recorded during 01/01/2004 - 31/12/2015, stratified by calendar year
Example of counts based on <u>two</u> criteria:
1) The total number of patients with a <u>medical diagnosis</u> of Type 2 diabetes mellitus recorded in CPRD GOLD or HES APC on or before 31/12/2005 OR
2) <u>Prescriptions</u> for anti-diabetic medication in CPRD GOLD (<i>note - provided in one code list</i>) on or before 31/12/2005. Patients must have at least 12 months of prior registration before their earliest event date.
Example of counts based on <u>three</u> criteria:
1) The total number of patients in CPRD GOLD or CPRD Aurum or HES APC with an incident <u>medical diagnosis</u> of Type 2 diabetes mellitus recorded in during 01/01/2004 - 31/12/2015 OR
2) Incident <u>prescription</u> of anti-diabetic medication (<i>note-provided in one code list</i>) documented during 01/01/2004 - 31/12/2015 AND
3) Have a <u>test record</u> for HbA1c recorded in CPRD Aurum (<i>note - test value not assessed</i>).

To request a free simple feasibility count you will need to prepare a relevant code list for the events of interest, in a tab-delimited text file. You can request the CPRD Code Browser (free of charge) which will allow you to search for medical and product codes. Please see the section '[Guidance on generating code lists](#)' for further information. For access to the CPRD Code Browser tool, please email enquiries@cprd.com.

What are the differences between the CPRD Aurum and CPRD GOLD databases?

The table below outlines some differences between the CPRD GOLD and the CPRD Aurum data that you may find useful if you are considering using the CPRD Aurum data alone, or in combination with CPRD GOLD. Further information can be found in the '[Points for consideration](#)' section.

Difference	Context	CPRD Aurum	CPRD GOLD	Further information
Medical coding	The NHS is moving to universal coding using SNOMED-CT	Clinical observations are recorded using a mixture of Read 2, SNOMED and local EMIS® codes.	Medical events are based on Read coding. SNOMED coding will be added to the GOLD database, and this will be mapped 1:1.	Advice on producing code lists in the CPRD Aurum (& CPRD GOLD) data is provided under Guidance on generating code lists .
Product coding			Product coding in CPRD GOLD uses Gemscript	Advice on producing code lists in the CPRD Aurum (& CPRD GOLD) data is provided under Guidance on generating code lists .
Test and value recording		Test and value results are recorded in the Observation table	Test and value results are recorded in the Additional Clinical Details and Test tables	Advice on finding measurements in CPRD Aurum can be found under Finding results of tests, investigations and other clinical measurements .
Vaccination recording		Vaccinations are recorded in the Observation table. There may also be records in the Drug Issue table.	Vaccinations are recorded in the Immunisation table. There may also be records in the Therapy table.	
Derived variables	CPRD offers several derived variables to facilitate research.	The CPRD Aurum database includes derived variables such as a derived death date and acceptable patient flag. There is no up-to-standard date	The CPRD GOLD database includes derived variables such as a derived death date, acceptable patient flag, and practice up-to-standard date.	Further information can be found under Derived variables .
Consultations		Events may be added to the patient record outside of the context of a consultation. Consultation identifiers may not be present.	All events are linked to a consultation by a consultation identifier.	

Points for consideration

Duplication of data between practices in the CPRD Aurum database

Duplication of patient data can occur when one practice is absorbed by another practice that also contributes (or goes on to contribute) to the CPRD Aurum database. This issue potentially affects all builds of the CPRD Aurum database to date.

When a practice is absorbed by another one it stops contributing data to the database, but its data remains in the database, in the same way as for all other practices that close or stop contributing data. Patients from the absorbed practice are assigned a new patient identifier in the absorbing practice, but carry across all their data to, and retain their original registration start date in, their new practice. This is different from what happens when individual patients move between two practices which contribute to the database. In that case previous data is also carried across but, because it is before their new registration date, studies can avoid duplication by excluding data recorded before each individual's registration date. Such people are effectively split into two separate individuals within analyses.

We have determined practices that appear likely to have merged into other contributing practices and suggest that you may want to exclude these from your studies, wherever possible. The practice identifiers are listed on the CPRD website in the most recent CPRD Aurum data specification document, and these are regularly updated.

These practices had more than 95% of their patients sharing combinations of registration start date, sex, and year of birth with patients in another practice within the same region. No other pairs of practices share more than 70% of values.

We intend to resolve this issue in future builds, but older builds will remain unchanged in order to allow for replication of previous analyses.

Duplication between the CPRD GOLD and CPRD Aurum databases

A number of GP practices that previously contributed data to CPRD GOLD, when using InPS Vision software, are now supported by EMIS software and have agreed to contribute data to CPRD Aurum. In this situation, CPRD will hold duplicate historical data for such practices in the CPRD GOLD and CPRD Aurum databases. If you are planning to use data from both databases for a study, CPRD can provide a migrators file to identify the overlapping practices and dates: VisionToEmisMigrators.txt. This migrators file is refreshed every month with the CPRD Aurum monthly build. Researchers can choose to remove migrating practices from either the CPRD GOLD data or the CPRD Aurum data, as required by the study. Please contact enquiries@cprd.com for access to this file, stating the monthly build you require.

Medical and Product dictionaries

Within EMIS clinical systems, healthcare professionals can record some observations using local codes, rather than Read or SNOMED CT codes. Where possible, local EMIS® codes have been mapped to SNOMED CT, but you may still find items in the medical dictionary that are not mapped to either Read or SNOMED codes. To add value to the CPRD Aurum product dictionary, CPRD has mapped it to the Dictionary of Medicines and Devices (DM+D).

For advice on producing code lists for CPRD Aurum, see the section [Guidance on generating code lists](#).

Recording of coded clinical information

Within EMIS clinical systems, healthcare professionals have a greater opportunity to use free text rather than coding to record clinical observations. CPRD does not receive free text due to information governance restrictions which may mean that there are systematic differences in the recording of observations between CPRD GOLD and CPRD Aurum. CPRD researchers have conducted a preliminary evaluation of CPRD GOLD and CPRD Aurum data and have found similar prevalence estimates for common conditions, including heart failure and chronic kidney disease, between the two databases. As CPRD increases its understanding of these issues, we will share our findings.

Identifying clinical measurements

In CPRD Aurum, clinical measurements such as blood pressure, height and weight are recorded in the observation table. Relevant measurements should be identified via a medical code list and presence of a value to filter observations. See '[Example research questions](#)' for further information.

How are referrals recorded?

Referral information is recorded in two separate tables: the Observation table contains details about the reason for the referral (as a medical code) and event date; the Referral table contains details about the source and target organisation, referral urgency and service type. The complete Referral record can be reconstructed by linking the Observation and Referral records using the observation identifier ('obsid') which is present in each table.

What is the Problem table?

GPs are able to assign 'problem' status to observations in the EMIS Web® software. This is a way of enabling GPs to view a patient's medical history by clinical issue rather than in chronological order.

For instance, classifying a patient's diabetes as a problem would allow them to link observations, such as diabetes medication reviews and blood tests, in order to better monitor their diabetes management. This table may contain valuable information in addition to the Observation table, but it is important to note that there could be variation in the way that different GPs use the 'problems' recording option. Problem information is recorded in two separate tables: the Observation table contains details about the nature of the issue (as a medical code) and event date; the Problem table contains further details including duration, clinical significance, whether the problem remains active. The complete Problem record can be reconstructed by linking the Observation and Referral records using the observation identifier ('obsid') which is present in each table.

Unique identifiers

CPRD Aurum contains a number of unique identifiers. These serve a dual purpose of allowing identification of unique records as well as unique linkage across dataset tables. We are aware that a small number of unique identifier records from data received from 2018 link to more than one record between the observations, consultations, drug issue and problem tables and in some instances do not have matching patient details. The issue has been resolved for the CPRD Aurum June 2024 release by removal of the affected records. For researchers using historic builds of CPRD Aurum we suggest you may want to exclude any such records which do not form unique linkages across tables.

National Data Opt-out and deceased/transferred-out patients prior to 1997

The national data opt-out (NDOO) was introduced on 25 May 2018 following recommendations from the National Data Guardian and indicates that a patient does not want their confidential patient information to be shared for purposes beyond their individual care across the health and care system in England. The mandatory implementation deadline of the NDOO was 31st July 2022.

The change involved treating patients as opted out nationally rather than locally. In terms of the opt-out process, any patient that had a type 2 opt-out recorded on a GP system on or before 11 October 2018 then had it automatically converted to a NDOO on implementation (in 2022).

Based on discussions with GP software suppliers, our understanding is that the technical implementation of the NDOO requires new format NHS numbers, which were introduced in 1995-96. As a result, primary care data for patients in England with an old format NHS number or a missing NHS number is no longer shared with CPRD and no longer appears in CPRD Aurum and GOLD databases. This particularly affects patients in England who died or transferred out of a participating

practice before they were issued a new NHS number, i.e. before 1997. CPRD are investigating options with NHS England for reinstating data for these patients in future. We are liaising with the NDOO Implementation team to verify this and explore whether there is a way to receive the data for these patients who have not otherwise registered an opt-out. In the meantime, researchers should be aware of this issue, and should consider excluding time before 1997 when defining a study period.

Linked data

Linkage of CPRD Aurum to all the standard patient-level linked datasets available for the CPRD GOLD database is now available. The process of linking CPRD Aurum to other datasets is the same as for CPRD GOLD. For more information: <https://cprd.com/linked-data>.

Guidance on applying to RDG

The RDG (Research Data Governance) application process for CPRD Aurum data is the same as for CPRD GOLD data. The electronic research applications portal (eRAP) offers the option to request CPRD Aurum data, CPRD GOLD data, or both. If you are new to using CPRD data we would advise you to speak to a researcher at CPRD about the feasibility of your proposed study, but this is not a requirement. It is expected that your RDG protocol details the considerations and limitations of using CPRD Aurum, for example if you are planning on conducting a study using both CPRD GOLD and CPRD Aurum data, you should consider potential differences in the databases that may impact your results. An understanding of these potential differences and the implications for your study conduct and findings should be demonstrated in your RDG protocol.

Guidance on generating code lists

The CPRD Aurum medical and product dictionaries are more complex than those for the CPRD GOLD database, and the CPRD GOLD dictionary cannot be considered a subset of the CPRD Aurum dictionary. The dictionaries are available through the CPRD Code Browser. The Code Browser can be requested by contacting enquiries@cprd.com. Please note that you will need to generate code lists separately for CPRD GOLD and CPRD Aurum (there is a separate set of dictionaries available for each source to facilitate this).

We advise that you use a combination of term and read code searches using the hierarchy.

At CPRD, our experience of working with both CPRD GOLD and CPRD Aurum databases has indicated that developing reusable search strategies for code list generation is preferable to maintaining static code lists. These strategies may combine searches of descriptor terms (for example, CKD or chronic kidney disease) and hierarchical classifications such as Read, to identify codes for inclusion or exclusion. The benefit of this strategy is that it can be re-used at a later date to update code lists. Further, it can be applied to generate code lists for both the CPRD GOLD and CPRD Aurum databases simultaneously rather than having to replicate a code list developed in one database. For large and complex code lists replication of an existing code list may be difficult.

CPRD strongly recommends that medcodes, prodcodes, SNOMED codes and any other long numeric identifiers are imported, stored, and processed as text rather than integers. In CPRD Aurum, unique identifiers in the Medical Dictionary (medcodeid) and the Product Dictionary (prodcoid) (as well as SNOMED codes) can be up to 18 digits in length. Standard software packages including R, Stata, SPSS, and Excel are unable to store integers of this magnitude without loss of precision. In other words, these software packages will retain incorrect approximations if these unique identifiers are stored as integers. The CPRD Aurum tools have been designed to overcome this limitation by importing, storing, and exporting text files.

Further advice is available from a CPRD researcher via enquiries@cprd.com.

Derived variables

The following derived variables are available in CPRD Aurum:

- CPRD death date (Patient table)
- Acceptable patient flag (Patient table)

The following derived variables are under development for later release. Fields for these variables are included in the data tables but will either not be populated or may contain data without a lookup:

- Up-to-standard date (Practice table)

The Up-to-standard date in the CPRD GOLD database is deemed as the date at which data in a practice is considered to be of a continuous high-quality that is fit for use in research. It is based on an assessment of recording of consultations and deaths within the practice. Until the Up-to-

standard date is available for CPRD Aurum, we advise that clients investigate patterns of practice-level recording for relevant observation records (including mortality) to explore capture of events throughout their study period. This is particularly relevant for studies looking back 15 years or more.

- CPRD consultation type (Consultation table)

Please see the Data Specification and Derived variables in CPRD GOLD and CPRD Aurum documents for further information, these are available at <https://cprd.com/primary-care>.

Example research questions

This section will be updated with additional problems and solutions as our understanding of the data increases. If there is a particular question on which you would like further information, please email enquiries@cprd.com and we will be happy to advise you.

Available results of tests, investigations, and other clinical measurements

Numeric results of tests, investigations, and other clinical measurements are available from the Observation table, as a combination of:

- A medical code [medcodeid] which describes the parameter being recorded (where the Medcodeid descriptive term is recorded in the EMISMedicalDictionary.txt file).
- A numeric value [value] providing the quantified result.
- A unit of measurement [numunitid] describing units of the measurement (where the actual unit description is recorded in the NumUnit.txt file)
- Optionally there may be two values to define the lower limit [numrangelow] and upper limit [numrangehigh] of the 'normal range' for the measurement.

Comparators, such as < and >, can be recorded for a “value” and are necessary for interpreting the numeric content of a “value” associated with a medcodeid. This comparator is recorded in EMIS systems but is not currently included in the data feed used to create CPRD Aurum.

We will explore whether the comparator variable can be incorporated into future builds of the Aurum database to facilitate the correct interpretation of the value data.

Table 1. Example of available numerical measurement result variables.

medcodeid	Term	value	Numunitid (description)	numrangelow	numrangehigh
100716012	Body mass index	30.610	kg/m2		
102928018	Platelet count	239.000	10*9/L	150.000	450.000
1227423014	Percentage eosinophils	4.000	%	0.000	7.000

145471000006115	Serum triglycerides level	2.000	mmol/L	0.500	2.300
275131000006114	Temperature	36.800	degrees C		
306221000000119	Serum alanine aminotransferase level	58.000	u/L	10.000	35.000
714931000006112	MCV - Mean corpuscular volume	87.000	fL	80.000	100.000
976481000006110	GFR (glomerular filtration rate) calculated by abbreviated Modification of Diet in Renal Disease Study Group calculation	76.000	ml/min/1.73m2	90.000	120.000
98461000006116	Total white cell count	6.100	10*9/L	3.700	11.000

Interpreting numeric test results in CPRD Aurum:

There is variability in the content and completeness of both the numeric "value" and the associated interpretative variables. In addition, interpretation of exact test results, or understanding which results fall above/below a set threshold may be impacted by lack of the comparator value. However, for researchers only interested in positive/negative results, or the presence of a test, this will be a less critical issue.

Useful investigations which may assist in assessing suitability of the 'value' data for research projects include:

- inspecting the degree of missingness in "value" and its associated interpretive variables.
- visualising the numeric distribution of results within "value" and comparing this to the available numeric measurement ranges.
- inspecting the proportion of results which fall within the expected ranges.
- inspecting the units available for interpretation within "numunitid".
- investigating factors such as testing guidance, available test Limits of Detection and Quantification and automated laboratory result reporting

Researchers are advised to assess the impact on their studies and consider conducting feasibility studies, where this data is required from CPRD Aurum, prior to submitting full study RDG reviews.

Additional documentation

General

Data resource profile: Clinical Practice Research Datalink (CPRD) Aurum

<https://doi.org/10.1093/ije/dyz034> (*Int J Epidemiol* – Open Access)

Accuracy of date of death recording in the CPRD GOLD database

<https://doi.org/10.1002/pds.4747> (*Pharmacoepidemiol Drug Saf* - Open Access)

NHS Digital: SNOMED CT resource

<https://digital.nhs.uk/snomed-ct>

Code list generation

Clinical code set engineering for reusing EHR data for research: A review.

<https://doi.org/10.1016/j.jbi.2017.04.010> (*J Biomed Inform* - Open Access)

Identifying clinical features in primary care electronic health record studies: methods for code list development

<https://doi.org/10.1136/bmjopen-2017-019637> (*BMJ Open* – Open Access)