CPRD Aurum
Frequently asked questions (FAQs)

Version 2.0

Date: 10th April 2019

Authors:

Helen Booth, Daniel Dedman, Achim Wolf (CPRD, UK)

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 By	Date	Reviewed By	Date
1.0		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 info	Helen Booth	01/03/2018		
1.4	Updated release figures	Helen Booth	04/04/2018		
1.5	Updated release figures & linkage info	Helen Booth	02/05/2018		
1.6	Updated release figures & linkage info	Helen Booth	02/07/2018		
2.0	Updated for CPRD Aurum tools	Achim Wolf	10/04/2019		

Contents

Introduction

What is CPRD Aurum?

Clinical Practice Research Datalink (CPRD) Aurum is a new 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 the EMIS Web® electronic patient record system from EMIS Health®, and the database has been named CPRD Aurum after the Latin word for 'gold'.

How does CPRD Aurum differ from the CPRD GOLD database?

CPRD Aurum contains data contributed by practices using EMIS Web® software, whilst CPRD GOLD holds data from a different software provider named 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. The basic content of the research tables in the CPRD Aurum database are similar to those in CPRD GOLD. The main differences are outlined in later sections of this document.

What is the population coverage of CPRD Aurum?

Monthly release notes are produced for the CPRD Aurum database and distributed to license holders. They are also available through the customer website. If you cannot access these please contact enquiries@cprd.com.

Currently, all practices contributing data to the CPRD Aurum database are in England, and CPRD is exploring options for the inclusion of further data from EMIS Web® practices in the devolved nations.

What does the CPRD Aurum database contain?

When an EMIS Web® practice agrees to contribute data to CPRD Aurum, 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. Where a practice has switched from another software system, such as Vision®, CPRD will receive all historical data, including information that was migrated from previously used software systems, in addition to prospectively added data. Consequently, there is some overlap between the CPRD GOLD and CPRD Aurum databases where practices have contributed via both software systems. Further information on dealing with this in studies intending to

use data from both databases is provided in the section on 'Duplication between the CPRD GOLD and CPRD Aurum databases'.

How can I access CPRD Aurum?

The ISAC application process for CPRD Aurum data will be the same as for CPRD GOLD, but applicants who have not previously used CPRD Aurum data should discuss their proposals with a CPRD researcher before submitting an application, to ensure an understanding of the data structure and implications for study design. Please see the section 'Guidance on applying to ISAC' for further advice on ISAC applications to use the CPRD Aurum data. CPRD online tools have been updated to enable access to both CPRD GOLD and CPRD Aurum.

What is the cost for accessing CPRD Aurum data?

Please contact enquiries@cprd.com for a quote for your specific study. CPRD Aurum access is also available through a multi-study licence.

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

Multi-study licence holders can access the database for feasibility purposes using the online tools. If you do not have access to the online tools CPRD can provide simple feasibility counts for CPRD Aurum and CPRD GOLD free-of-charge.

- 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).
- Examples of simple feasibility counts based on 1-3 criteria are outlined in the table below.
- 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 include:

Examples of counts based on one criterion
The total number of metformin prescriptions recorded in CPRD Aurum data during 01/01/2004 - 31/12/2015, stratified by calendar year.
Examples of counts based on two criteria
Separate counts of the total number of patients with at least one prescription for metformin or sulfonylureas recorded in CPRD Aurum data during 01/01/2004 - 31/12/2015, stratified by calendar year of first prescription. Two separate counts will be provided.
Examples of counts based on three criteria
Separate counts of the total number of patients with at least one prescription for metformin, sulfonylureas or thiazolidinediones recorded in CPRD Aurum data during 01/01/2004 - 31/12/2015, stratified by calendar year of first prescription. Three separate counts will be provided.

To request a free simple feasibility count you will need to send CPRD a code list of medical events or prescriptions that can be used with the CPRD Aurum dictionaries, in tab delimited text files. Please see the section 'Guidance on generating code lists' for further information. We are happy to advise you on this process. Clients can request more sophisticated feasibility counts and are advised to discuss their needs with a CPRD researcher so that a service quote can be provided if relevant. Please email enquiries@cprd.com to discuss your needs further.

What are the major 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. Differences that are temporary have been highlighted. 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	Source coding in CPRD Aurum uses a mixture of Read 2, SNOMED and local EMIS® codes.	Source-coding in CPRD GOLD is 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 here.
Product coding		Product coding in CPRD Aurum uses DM+D	Product coding in CPRD GOLD uses Gemscript	Advice on producing code lists in the CPRD Aurum (& CPRD GOLD) data is provided here.
Test and value recording*		In CPRD Aurum test and value results are	In CPRD GOLD test and value results are	Advice on finding measurements in

		recorded in the Observation table	recorded in the Additional Clinical Details and Test tables	CPRD Aurum can be found here.
Vaccination recording		In CPRD Aurum vaccinations are recorded in the Observation table	In CPRD GOLD vaccinations are recorded in a separate immunisation table	
Derived variables**	CPRD offers a number of derived variables to facilitate research in the CPRD GOLD database, such as a derived death date, acceptable flag, up-to-standard date	The current CPRD Aurum database release version includes the acceptability flag and derived death date variables. The up-to-standard date is undergoing development.	No change	Some advice is provided on variables to be introduced, and alternative approaches that CPRD users may wish to adopt in the interim here.
Consultations		Clinical observation recording may be added outside of the context of a consultation. Consultation identifiers may not be present	Observations are all linked to a consultation by a consultation identifier	

* See 'Points for consideration' for further information

** See 'Derived variables' for further information

Points for consideration

Duplication between the CPRD GOLD and CPRD Aurum databases

A number of GP practices that previously contributed data to CPRD GOLD are now supported by EMIS Web® 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 bridging file to identify the overlapping practices and dates.

Medical and Drug dictionaries

EMIS Web® software enables clinicians to 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 drug 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

EMIS Web® software offers greater opportunity for GPs 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 data sourced from EMIS Web® as compared to data sourced from Vision® for research 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 with clients. However, this is a difference that you should be aware of when preparing definitions and code lists to be used with the CPRD Aurum data.

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 problem (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.

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.

Set 16 linkages are available for 232 CPRD Aurum practices. For set 17 (due to be released May 2019), there will be linkage for 800 CPRD Aurum practices. There is a lag between transfer of linkage identifiers from EMIS® to NHS Digital and receipt of the linked data at CPRD. Consequently, while the number of practices in CPRD Aurum is growing rapidly, the number of practices with linked data will be lower than the total number in the CPRD Aurum database.

Guidance on applying to ISAC

The ISAC application process for the CPRD Aurum data will be the same as for the CPRD GOLD data. The form offers a tick-box option to request CPRD Aurum data. If you are considering a study using CPRD Aurum we would expect you to speak to a senior researcher at CPRD for advice on the feasibility of your proposed study. Enquiries can be sent to enquiries@cprd.com and will be directed

to the appropriate person. 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 ISAC protocol.

Guidance on generating code lists

The CPRD Aurum dictionaries are more complex than those for the CPRD GOLD database, as described previously. The dictionaries are available through the CPRD code browser. The CPRD code browser and a user guide can be requested by contacting enquiries@cprd.com. If you are already using the code browser to search the CPRD GOLD dictionaries you will still need to contact us to download the browser containing the CPRD Aurum dictionaries.

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.

Medcodes, prodcodes, and any SNOMED codes should always be stored as 'string' (text) and not as integers. In CPRD Aurum, unique identifiers in the Medical Dictionary (medcodeid) and the Product Dictionary (prodcodeid) (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 now available in CPRD Aurum:

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

The following derived variables are still being developed and will be added in later release versions:

- Up-to-standard date (Practice table)
- CPRD consultation type (Consultation table)

Fields for these variables are included in the data tables, but will either not be populated or may contain data without a lookup. Please see the Data Specification document for further information.

Acceptable patient flag

Patients are classified as not acceptable if they meet any of the following criteria:

- Year of birth is empty
- Current registration date is empty
- Current registration date is greater than the practice's last collection date
- Current registration date is less than or equal to 01/01/1900
- Current registration date is equal to or greater than the registration end date
- Current registration date is prior to the birth year
- Gender other than male, female or indeterminate
- Age is greater than 115 at end of follow-up (based on registration end date, death or last collection date)
- All recorded health care episodes have empty event dates
- All recorded health care episodes have invalid events dates (less than or equal to 01/01/1900 or greater than last collection date)
- All recorded health care episodes have dates before the birth year
- Patients are not permanently registered

This algorithm is the same as that applied to the CPRD GOLD data with the exception of critera utilising 'first registration date' and 'transfer-out reason', which are available in CPRD GOLD data but not in CPRD Aurum.

Derived death date

Information on date of death is included in the source data used for CPRD Aurum (emis_ddate), but this may not always correspond to the date of occurrence. For instance, it may reflect the date of notification of the death to the GP, or when the deceased patients' registration record was updated. CPRD therefore provide a more realistic estimate of the date of death (cprd_ddate), based on an algorithm which uses additional information in the patient record. The algorithm is similar to that used in CPRD GOLD, and the resulting estimate should be sufficiently accurate for purposes such as

censoring follow-up time. For studies where accurate date, and/or cause of death are important, use of linked ONS mortality data is recommended.

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.

Finding results of tests, investigations and other clinical measurements

Numeric results of tests, investigation and other clinical measurements are recorded in the Observation table, as a combination of:

- A medical code [medicalcodeid] which describes the parameter being record. The text description [term] associated with the code can be obtained from the medical dictionary.
- A numeric value [value]
- A unit of measurement [numunitid]
- Optionally there may be two values to define the lower limit [numrangelow] and upper limit [numrangehigh] of the 'normal range' for the measurement.

Examples of numeric results for tests, investigations and clinical measurements:

1. Initially, the medical dictionary should be searched for Read terms that could be used to record the measurement of interest. For blood pressure, a search using the terms ‘systolic blood pressure’ and ‘diastolic blood pressure’ can be used to produce a code list to identify relevant observations.
2. The codelist can then be applied to the Observation table to identify observations that include measurements for blood pressure in the ‘*value’* field.
3. The identified observations can then be cleaned by checking whether a value has been recorded, and using the *‘numunitid’* to check the measurements have the appropriate unit (mmHg).

Additional documentation

The following documents may be useful for background information.

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