

CPRD Aurum Data Specification

Version 2.3

Date: 23 August 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.

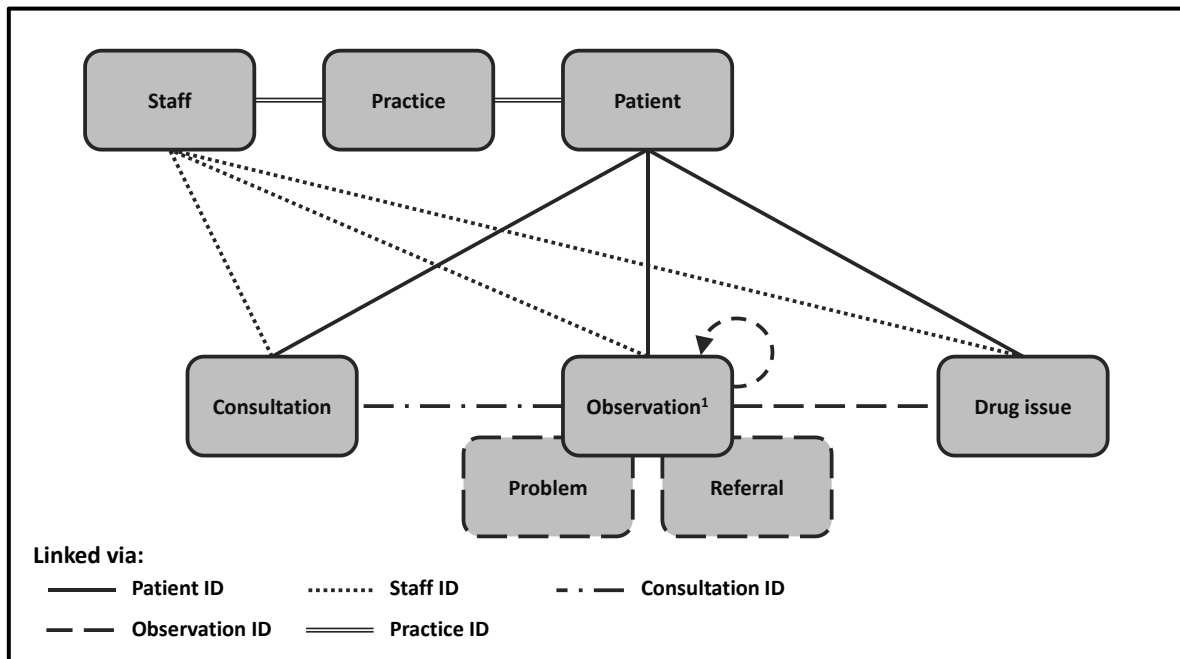
| Version | Summary of Change | Prepared By | Date | Reviewed By | Date |
|---------|---|--------------|------------|---------------------------|------------|
| 1.0 | | Helen Booth | 06/12/2017 | | |
| 1.1 | Update fields | Helen Booth | 11/01/2018 | | |
| 1.2 | Update fields | Helen Booth | 29/01/2018 | | |
| 2.0 | Update for release of tools | Achim Wolf | 10/04/2019 | | |
| 2.1 | Updates to Drug Issue table from DTT feedback | Helen Booth | 23/04/2019 | | |
| 2.2 | Reviewed and updated all data types | Helen Booth, | 01/05/2019 | Jenni Chapman, Dan Dedman | 02/05/2019 |
| 2.3 | Reviewed and updated based on client feedback | Dan Dedman | 07/08/2019 | Achim Wolf | 09/08/2019 |

Data Format

The data are available to researchers as eight files in text format as listed below, with a graphical representation below.

1. The **Patient** file (Patient/NNN.txt) contains basic patient demographics and patient registration details for the patients.
2. The **Practice** file (Practice001.txt) contains details of each practice, including the practice identifier, practice region, and the last collection date.
3. The **Staff** file (Staff/NNN.txt) contains practice staff details for each staff member, including job category.
4. The **Consultation** file (Consultation/NNN.txt) contains information relating to the type of consultation as entered by the GP (e.g. telephone, home visit, practice visit). Some consultations are linked to observations that occur during the consultation via the consultation identifier (consid).
5. The **Observation** file (Observation/NNN.txt) contains the medical history data entered on the GP system including symptoms, clinical measurements, laboratory test results, and diagnoses, as well as demographic information recorded as a clinical code (e.g. patient ethnicity). Observations that occur during a consultation can be linked via the consultation identifier. CPRD Aurum data are structured in a long format (multiple rows per subject), and observations can be linked to a parent observation. For example, measurements of systolic and diastolic blood pressure will be grouped together via a parent observation for blood pressure measurement.
 - a. The **Referral** file (Referral/NNN.txt) contains referral details recorded on the GP system. Data in the referral file are linked to the observation file and contain 'add-on' data for referral-type observations. These files contain information involving both inbound and outbound patient referrals to or from external care centres (normally to secondary care locations such as hospitals for inpatient or outpatient care). To obtain the full referral record (including reason for the referral and date), referrals should be linked to the Observation file using the observation identifier (obsid).
 - b. The **Problem** file (Problem/NNN.txt) contains details of the patient's medical history that have been defined by the GP as a 'problem'. Data in the problem file are linked to the observation file and contain 'add-on' data for problem-type observations. Information on identifying associated problems, the significance of the problem and its expected duration can be found in this table. GPs may use 'problems' to manage chronic conditions as it would allow them to group clinical events (including drug prescriptions, measurements, symptom recording) by problem rather than chronologically. To obtain the full problem record (including the clinical code for the problem), problems should be linked to the Observation file using the observation identifier (obsid).
6. The **Drug issue** file (DrugIssue/NNN.txt) contains details of all prescriptions on the GP system. This file contains data relating to all prescriptions (for drugs and appliances) issued by the GP. Some prescriptions are linked to problem-type observations via the Observation file, using the observation identifier (obsid).

CPRD Aurum dataset structure



¹Includes symptoms, diagnoses, immunisations, tests, and lifestyle factors. Note: The problem and referral tables contain add-on information for certain types of observations. Some consultations are linked to observations. Some drug issues are linked to problem-type observations.

Data dictionaries

The CPRD Aurum dictionaries will be provided as text files that can be imported into standard statistical software to enable code searching. The dictionaries are also 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 latest browser containing the CPRD Aurum dictionaries. For advice on producing code lists please refer to the CPRD Aurum FAQ documentation.

- I. The **Medical** Dictionary contains information on all medical history observations that have been recorded in EMIS Web[®]. Observations are coded using a combination of SNOMED, Read and local EMIS[®] codes. Further information is provided in later sections of this document.
- II. The **Product** Dictionary contains information on drug and appliance prescriptions recorded in EMIS Web[®]. This information is coded using the Dictionary of Medicines and Devices (DM+D). Further information is provided in later sections of this document.

**** IMPORTANT ****

CPRD strongly recommends that medcodes, prodcodes, SNOMED codes and any other long numeric identifiers are stored and processed as text rather than integers. In CPRD Aurum, unique identifiers such as these can be up to 19 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.

Fields where this recommendation applies are indicated in the table specifications below as having a Field Type of 'TEXT', and a Format which includes numeric characters only.

Field descriptions

Full descriptions of fields in each data file are provided in the tables below. Most files can be linked using the encrypted patient identifier (patid). The last five digits of the patient identifier (patid) denote the identifier of the practice (pracid) that the patient belongs to. The mapping column lists lookup files with further information on decoding numerical values.

1. Patient

| Column name | Field name | Description | Mapping | Type | Format |
|--------------------------|----------------|---|----------------------------|---------|-----------------------------|
| Patient identifier | patid | Encrypted unique identifier given to a patient in CPRD Aurum. The patient identifier is unique to CPRD Aurum, and may represent a different patient in the CPRD GOLD database. This is the primary key for this table. The last 5 characters will be same as the CPRD practice identifier | | TEXT | 6-19 numeric characters |
| CPRD practice identifier | pracid | Encrypted unique identifier given to a practice in CPRD Aurum | Link Practice table | INTEGER | 5 |
| Usual GP | usualgpstaffid | The GP that the patient is nominally registered with. To be used with the Staff table for reference | Lookup: Staff table | TEXT | Up to 10 numeric characters |
| Gender | gender | Patient's gender | Lookup: Gender.txt | INTEGER | 3 |
| Year of birth | yob | Patient's year of birth. This is actual year of birth e.g. 1984. | | INTEGER | 4 |
| Month of birth | mob | Patient's month of birth (for those aged under 16). | | INTEGER | 2 |
| Date of death | emis_ddate | Date of death as recorded in the EMIS® software. Researchers are advised to treat the <i>emis_ddate</i> with caution and consider using the <i>cprd_ddate</i> variable below. | | DATE | DD/MM/YYYY |
| Registration start date | regstartdate | The date that the patient registered with the CPRD contributing practice. Most recent date the patient is recorded as having registered at the practice. If a patient deregistered for a period of time and returned, the return date would be recorded. | | DATE | DD/MM/YYYY |
| Patient type | patienttypeid | The category that the patient has been assigned to e.g. private, regular, temporary. | Lookup: PatientType.txt | INTEGER | 5 |
| Registration end date | regenddate | Date the patient's registration at the practice ended. This may represent a transfer-out date or death date. | | DATE | DD/MM/YYYY |
| Acceptable flag | acceptable | Flag to indicate whether the patient has met certain quality standards: 1 =acceptable, 0 = unacceptable | | INTEGER | 1 |
| CPRD death date | cprd_ddate | Estimated date of death of patient – derived using a CPRD algorithm | | DATE | DD/MM/YYYY |

2. Practice

| <i>Column name</i> | <i>Field name</i> | <i>Description</i> | <i>Mapping</i> | <i>Type</i> | <i>Format</i> |
|--------------------------|-------------------|---|-----------------------|-------------|---------------|
| CPRD practice identifier | pracid | Encrypted unique identifier given to a practice in CPRD Aurum. This is the primary key for this table. | | INTEGER | 5 |
| Last Collection Date | lcd | Date of the most recent CPRD data collection for the practice. | | DATE | DD/MM/YYYY |
| Up-to-standard date | uts | This field will be empty while a CPRD algorithm is developed for the CPRD Aurum database. Once filled, it will represent the date at which the practice data is deemed to be of research quality. | | DATE | DD/MM/YYYY |
| Region | region | Value to indicate where in the UK the practice is based. The region denotes the Strategic Health Authority for English practices. | Lookup: Region.txt | INTEGER | 5 |

3. Staff

| <i>Column name</i> | <i>Field name</i> | <i>Description</i> | <i>Mapping</i> | <i>Type</i> | <i>Format</i> |
|--------------------------|-------------------|---|----------------------------|-------------|-----------------------------|
| Staff identifier | staffid | Encrypted unique identifier given to the practice staff member in CPRD Aurum. This is the primary key for this table. | Lookup: Staff table | TEXT | Up to 10 numeric characters |
| CPRD practice identifier | pracid | Encrypted unique identifier given to a practice in CPRD Aurum | Link Practice table | INTEGER | 5 |
| Job category | jobcatid | Job category of the staff member who created the event | Lookup JobCat.txt | INTEGER | 5 |

4. Consultation

| <i>Column name</i> | <i>Field name</i> | <i>Description</i> | <i>Mapping</i> | <i>Type</i> | <i>Format</i> |
|--------------------------------------|-------------------|--|--|-------------|-----------------------------|
| Patient identifier | patid | Encrypted unique identifier given to a patient in CPRD Aurum. The patient identifier is unique to CPRD Aurum, and may represent a different patient in the CPRD GOLD database. | Link Patient table | TEXT | 6-19 numeric characters |
| Consultation identifier | consid | Unique identifier given to the consultation. This is the primary key for this table. | | TEXT | Up to 19 numeric characters |
| CPRD Practice identifier | pracid | Encrypted unique identifier given to a practice in CPRD Aurum | Link Practice table | INTEGER | 5 |
| Event date | consdate | Date associated with the event | | DATE | DD/MM/YYYY |
| Entered date | enterdate | Date the event was entered into the practice system | | DATE | DD/MM/YYYY |
| Staff identifier | staffid | Encrypted unique identifier given to the practice staff member who took the consultation in CPRD Aurum | Link Staff table | TEXT | Up to 10 numeric characters |
| EMIS® consultation source identifier | conssourceid | Source of the consultation as recorded in the EMIS® software. Many of the fields in this column are generic e.g. GP surgery, telephone, but many contain identifiers such as practice name. Until a process for anonymising and cleaning this variable is established a lookup will not be provided. | Lookup: ConSource.txt [not included in initial release] | TEXT | 19 |
| CPRD consultation source identifier | cprdconstype | Source of the consultation e.g. telephone, main surgery. This lookup will be a compressed version of the 'conssourceid' i.e. where practice names have been recorded they will be compressed to 'GP practice'. | Lookup: cprdconstype.txt [not included in initial release] | INTEGER | 3 |
| Consultation source code identifier | consmedcodeid | Source of the consultation from EMIS® software. This is a medical code that can be used with the medical dictionary. It may contain information similar to the consultation source identifiers, but is available for use now. Some of the codes may not be interpretable e.g. Awaiting clinical code migration to EMIS Web®. | Medical dictionary. Maps to medcodeid | TEXT | 6-18 numeric characters |

5. Observation

| Column name | Field name | Description | Mapping | Type | Format |
|--------------------------------|--------------|---|--------------------------------|---------|-----------------------------|
| Patient identifier | patid | Encrypted unique identifier given to a patient in CPRD Aurum. The patient identifier is unique to CPRD Aurum, and may represent a different patient in the CPRD GOLD database. | Link Patient table | TEXT | 6-19 numeric characters |
| Consultation identifier | consid | Linked consultation identifier. In EMIS Web® it is not necessary to enter observations within a consultation, so this identifier may be missing. | Link Consultation table | TEXT | Up to 19 numeric characters |
| CPRD Practice identifier | pracid | Encrypted unique identifier given to a practice in CPRD Aurum | Link Practice table | INTEGER | 5 |
| Observation identifier | obsid | Unique identifier given to the observation. This is the primary key for this table. | | TEXT | Up to 19 numeric characters |
| Event date | obsdate | Date associated with the event | | DATE | DD/MM/YYYY |
| Entered date | enterdate | Date the event was entered into EMIS Web® | | DATE | DD/MM/YYYY |
| Staff identifier | staffid | Encrypted unique identifier given to the practice staff member who took the consultation in CPRD Aurum | Link Staff table | TEXT | Up to 10 numeric characters |
| Parent observation identifier | parentobsid | Observation identifier (obsid) that is the parent to the observation. This enables grouping of multiple observations, such as systolic and diastolic blood pressure, or blood test results. | Link Observation table | TEXT | Up to 19 numeric characters |
| Medical code | medcodeid | CPRD unique code for the medical term selected by the GP | Lookup: Medical dictionary | TEXT | 6-18 numeric characters |
| Value | value | Measurement or test value | | NUMERIC | 19.3 |
| Numeric unit identifier | numunitid | Unit of measurement | Lookup: NumUnit.txt | INTEGER | 10 |
| Observation type identifier | obstypeid | Type of observation (allergy, family history, observation) | Lookup: ObsType.txt | INTEGER | 5 |
| Numeric range low | numrangelow | Value representing the low boundary of the normal range for this measurement | | NUMERIC | 19.3 |
| Numeric range high | numrangehigh | Value representing the high boundary of the normal range for this measurement | | NUMERIC | 19.3 |
| Problem observation identifier | probobsid | Observation identifier (obsid) of any problem that an observation is associated with. An example of this might be an overarching condition that the current observation is associated with such as 'wheezing' with the problem observation identifier that links to an observation of 'asthma', that in turn contains information in the problem table. | Link Observation table | TEXT | Up to 19 numeric characters |

5a. Referral

| <i>Column name</i> | <i>Field name</i> | <i>Description</i> | <i>Mapping</i> | <i>Type</i> | <i>Format</i> |
|---|-------------------|--|---|-------------|-----------------------------|
| Patient identifier | patid | Encrypted unique identifier given to a patient in CPRD Aurum. The patient identifier is unique to CPRD Aurum, and may represent a different patient in the CPRD GOLD database. | Link Patient table | TEXT | 6-19 numeric characters |
| Observation identifier | obsid | Unique identifier given to the observation | Link Observation table | TEXT | Up to 19 numeric characters |
| CPRD Practice identifier | pracid | Encrypted unique identifier given to a practice in CPRD Aurum | Link Practice table | INTEGER | 5 |
| Referral source organisation identifier | refsourceorgid | Source organisation of the referral. Organisations are identified by an ID number and each organisation has a type (e.g. hospital inpatient department, community clinic). Both the organisation table and the OrgType lookup are required. The lookups are undergoing anonymisation work. | Lookups: Organisation.txt and OrgType.txt [not included in initial release] | INTEGER | 10 |
| Referral target organisation identifier | reftargetorgid | Source organisation of the referral. Organisations are identified by an ID number and each organisation has a type (e.g. hospital inpatient department, community clinic). Both the organisation table and the OrgType lookup are required. The lookups are undergoing anonymisation work. | Lookups: Organisation.txt and OrgType.txt [not included in initial release] | INTEGER | 10 |
| Referral urgency identifier | refurgencyid | Urgency of the referral e.g. routine, urgent, dated | Lookup: RefUrgency.txt | INTEGER | 1 |
| Referral service type identifier | refservicetypeid | Type of service the referral relates to e.g. assessment, management, investigation | Lookup: RefServiceType.txt | INTEGER | 2 |
| Referral mode identifier | refmodeid | Mode by which the referral was made e.g. telephone, written | Lookup: RefMode.txt | INTEGER | 1 |

5b. Problem

| <i>Column name</i> | <i>Field name</i> | <i>Description</i> | <i>Mapping</i> | <i>Type</i> | <i>Format</i> |
|--|-------------------|---|-------------------------------|-------------|-----------------------------|
| Patient identifier | patid | Encrypted unique identifier given to a patient in CPRD Aurum. The patient identifier is unique to CPRD Aurum, and may represent a different patient in the CPRD GOLD database. | Link Patient table | TEXT | 6-19 numeric characters |
| Observation identifier | obsid | Unique identifier given to the observation. | Link Observation table | TEXT | Up to 19 numeric characters |
| CPRD Practice identifier | pracid | Encrypted unique identifier given to a practice in CPRD Aurum | Link Practice table | INTEGER | 5 |
| Parent problem observation identifier | parentprobobsid | Observation identifier for the parent observation of the problem. This can be used to group problems via the Observation table. | Link Observation table | TEXT | Up to 19 numeric characters |
| Problem end date | probenddate | Date the problem ended | | DATE | DD/MM/YYYY |
| Expected duration | expduration | Expected duration of the problem in days | | INTEGER | 5 |
| Last review date | lastrevdate | Date the problem was last reviewed | | DATE | DD/MM/YYYY |
| Last review staff identifier | lastrevstaffid | Staff member who last reviewed the problem | Link Staff table | TEXT | Up to 10 numeric characters |
| Parent problem relationship identifier | parentprobreleid | Description of the relationship of the problem to another problem e.g. the problem may have evolved or been merged with another problem as the problem, or the GP's understanding of the problem, changes | Lookup ParentProblemRel.txt | INTEGER | 5 |
| Problem status identifier | probstatusid | Status of the problem (active, past) | Lookup: ProbStatus.txt | INTEGER | 5 |
| Significance | signid | Significance of the problem (minor, significant) | Lookup: Sign.txt | INTEGER | 5 |

6. Drug issue

| <i>Column name</i> | <i>Field name</i> | <i>Description</i> | <i>Mapping</i> | <i>Type</i> | <i>Format</i> |
|--------------------------------|-------------------|--|---|-------------|-----------------------------|
| Patient identifier | patid | Encrypted unique identifier given to a patient in CPRD Aurum. The patient identifier is unique to CPRD Aurum, and may represent a different patient in the CPRD GOLD database. | Link Patient table | TEXT | 6-19 numeric characters |
| Issue record identifier | issueid | Unique identifier given to the issue record. This is the primary key for this table. | | INTEGER | 10 |
| CPRD Practice identifier | pracid | Encrypted unique identifier given to a practice in CPRD Aurum | Link Practice table | INTEGER | 5 |
| Problem observation identifier | probobsid | Unique identifier for the observation that links to a problem under which this prescription was issued. This refers to an 'obsid' in the Observation table which, in turn, can be linked to a record in the Problem table using 'obsid'. | Link Observation and Problem tables | TEXT | Up to 19 numeric characters |
| Drug record identifier | drugrecid | Unique identifier to a drug template record, which is not currently for release. At present this may be used to group repeat prescriptions from the same template. | | INTEGER | Up to 19 numeric characters |
| Event date | issuedate | Date associated with the event | | DATE | DD/MM/YYYY |
| Entered date | enterdate | Date the event was entered into EMIS Web® | | DATE | DD/MM/YYYY |
| Staff identifier | staffid | Encrypted unique identifier given to the practice staff member issued the treatment in CPRD Aurum | Link Staff table | TEXT | Up to 10 numeric characters |
| Drug code identifier | prodcodeid | Unique CPRD code for the treatment selected by the GP | Lookup: Product dictionary | TEXT | 6-18 numeric characters |
| Dosage identifier | dosageid | Identifier that allows dosage information on the event to be retrieved. Not included in first release | Lookup: common_dosages.txt | TEXT | |
| Quantity | quantity | Total quantity entered by the GP for the prescribed treatment | | DECIMAL | 9.3 ¹ |
| Quantity unit identifier | quantunitid | Unit of the treatment (capsule, tablet) | Lookup: QuantUnit.txt | INTEGER | 2 |
| Course duration in days | duration | Duration of the treatment in days | | INTEGER | 10 |
| Estimated NHS cost | estnhscost | Estimated cost of the treatment to the NHS | | DECIMAL | 10.4 ¹ |

¹ The number before the decimal point gives the precision i.e. the total number of digits. The number after the decimal point denotes the scale i.e. the maximum number of decimal places

I. Medical dictionary

| <i>Column name</i> | <i>Description</i> | <i>Mapping</i> | <i>Type</i> | <i>Format</i> |
|--------------------|---|----------------------------|-------------|-----------------------------|
| medcodeid | CPRD code to describe the observation. Links to the observation table | | TEXT | 6-18 numeric characters |
| term | Description of the observation associated with the codeid | | TEXT | 255 characters |
| originalreadcode | The original Read code text as provided in the EMIS® dictionary (contains codes which are not valid Read codes) | | TEXT | 100 characters |
| cleansedreadcode | CPRD-cleaned and validated version of the originalreadcode | | TEXT | 7 characters |
| snomedctconceptid | The SNOMED CT concept identifier associated with the observation | | TEXT | Up to 19 numeric characters |
| snomedctdescid | The SNOMED CT description identifier associated with the observation | | TEXT | Up to 19 numeric characters |
| emiscodecatid | Category identifier in EMIS® that describes the observation | Lookup: EMIScodecat.txt | INTEGER | 2 |
| release | Currently empty | | | |

II. Product dictionary

| <i>Column name</i> | <i>Description</i> | <i>Mapping</i> | <i>Type</i> | <i>Format</i> |
|-----------------------|---|----------------|-------------|-----------------------------|
| prodcodeid | CPRD code to describe the treatment. Links to the Drug Issue table | | TEXT | 6-18 numeric characters |
| termfromemis | Description of the treatment provided by EMIS® associated with the prodcodeid | | TEXT | 255 characters |
| dmdcode | The DM+D code associated with the treatment | | TEXT | Up to 19 numeric characters |
| productname | Name of the product | | TEXT | Up to 999 characters |
| formulation | Formulation of the product | | TEXT | Up to 999 characters |
| routeofadministration | Route of administration for the product | | TEXT | Up to 999 characters |
| drugsubstancename | Active ingredient(s) included in the product. For combination therapies, each component is listed, separated by / | | TEXT | Up to 999 characters |
| substancestrength | Strength of each active ingredient listed in the drugsubstancename column, including units. Separated by / if more than 1 | | TEXT | Up to 999 characters |
| bnfchapter | BNF chapter to which the product belongs | | TEXT | Up to 999 characters |
| release | Currently empty | | | |