**CPRD Aurum download December 2020: Importing into MySQL**

1. Tables available

a) Data tables

|  |  |  |
| --- | --- | --- |
| **Table** | **Description** | **Required (now)?** |
| Consultation | * Each Consultation has Consultation ID * Details on type of consultation * Includes Staff and Patient IDs | No |
| Drug issue | * Each Issue has Issue ID * Prescriptions * Includes Staff and Patient IDs | Yes |
| Observation | * Each Observation has Observation ID * Clinical codes (diagnoses, medical history, also demographic info) * Includes tests and results (not in a separate table as in CPRD Gold) * Can have a parent observation e.g. diastolic and systolic BP with BP as the parent – but from looking at subset of June 2020 CPRD Aurum download, most (75%) HbA1c tests do not have a parent observation – not much use. * Includes Staff ID and Consultation ID to link to those tables | Yes |
| Patient | * Each Patient has Patient ID * Basic demographic info * Registration/deregistration dates (for current registration; with Practice ID – link to Practice table) * Death date if applicable | Yes |
| Practice | * Each Practice has Practice ID * Practice region * Date of most recent data collection | Yes |
| Problem | * There is an option available on EMIS for doctor to link clinical codes and prescriptions to one ‘problem’ (an observation) e.g. a chronic illness like diabetes * This doesn’t appear to be utilised very much from looking at subset of June 2020 download. | No |
| Referral | * Referrals to and from external care centres (mostly to secondary care) * Includes Observation ID | No |
| Staff | * Each Staff has Staff ID * Job category for each Staff ID * Includes Practice ID | No |

b) Lookup tables

|  |  |  |
| --- | --- | --- |
| **Table** | **Description** | **Required (now)?** |
| Medical Dictionary | Medical codes used in Observation table | Yes |
| Product Dictionary | Product codes used in Drug issue table | Yes |
| common\_dosages | Dosage IDs used in Drug issue table | Yes |
| ConsSource | Where consultation type came from e.g. clinical notes, used in Consultation table | No |
| EmisCodeCat | Type of medical code e.g. test, used in Medical Dictionary | Yes\* |
| Gender | Used in Patient table | Yes |
| JobCat | Used in Staff table | No |
| NumUnit | Unit of measurement, used in Observation table | Yes |
| ObsType | Type of observation e.g. allergy, used in Observation table | Yes\* |
| OrgType | Source organisation of referral, used in Referral table | No |
| ParentProbRel | Relationship between two problems, used in Problem table | No |
| PatientType | Type of patient e.g. private, used in Patient table | Yes |
| ProbStatus | Problem status e.g. active, past, used in Problem table | No |
| QuantUnit | Unit of the treatment (capsule, tablet), used in Drug issue table | Yes |
| RefMode | How referral was made e.g. phone, used in Referral table | No |
| RefServiceType | Type of service the referral relates to e.g. assessment, used in Referral table | No |
| RefUrgency | How urgent referral is, used in Referral table | No |
| Region | Region (Strategic Health Authority) of practice, used in Practice table | Yes |
| Sign | Significance of the problem, used in Problem table | No |
| VisionToEmisMigrators | Practice IDs for practices previously in CPRD Gold and emigration to EMIS date | No |

\*extra info:

a) EMIS code categories: when I used a subset of the June 2020 CPRD Aurum download to look at HbA1c codes, I found all of the HbA1c medcodes which I identified in the Medical Dictionary had an EMIS code category of ‘Haematology’ (28 out of 35 HbA1c medcodes), ‘Unset’ (4 out of 35), or ‘Biochemistry’ (3 out of 35). Having these didn’t add anything but is a good sanity check.

b) Observation types: when I used a subset of the June 2020 CPRD Aurum download to look at HbA1c codes, I found all of the Observations with the HbA1c medcodes which I had identified were of Observation type ‘Investigation’ or ‘Value’. This could vary for a specific medcode. Again, having these didn’t add anything but is a good sanity check.

2. Tables to import and size

a) Data tables

|  |  |  |  |
| --- | --- | --- | --- |
| **Table** | **Field count** | **Line count (male/female counts don’t include 1st line)** | **Size of raw .txt files (GB)** |
| Drug issue | 14 | Male: 559,253,817  Female: 591,287,355  Total +1: 1,150,541,173 | Male: 95.6  Female: 101.0  Total: 196.7 |
| Observation | 15 | Male: **1,077,941,564**\*  Female: 1,097,085,653  Total +1: 2,175,027,218 | Male: 102.9  Female: 104.6  Total: 206.5 |
| Patient | 12 | Male: 765,277  Female: 716,017  Total +1: 1,481,295 | Male: 0.05  Female: 0.04  Total: 0.09 |
| Practice | 4 | NB: male and female files are duplicates (confirmed)  Total: 1442 | NB: male and female files are duplicates (confirmed)  Total: <0.01 |

\*See note in 3. a) ii) re. duplicate row in males Observations data

b) Lookup tables (May 2020 versions)

|  |  |  |  |
| --- | --- | --- | --- |
| **Table** | **Field count** | **Line count (-header)** | **Size of raw .txt files (MB)** |
| Medical Dictionary | 8 | 240,486 | 21.6 |
| Product Dictionary | 10 | 112,148 | 22.1 |
| common\_dosages | 11 | 72,749 | 8.4 |
| EmisCodeCat | 2 | 48 | <0.01 |
| Gender | 2 | 4 | <0.01 |
| NumUnit | 2 | 18,436 | 0.4 |
| ObsType | 2 | 11 | <0.01 |
| PatientType | 2 | 32 | <0.01 |
| QuantUnit | 2 | 97 | <0.01 |
| Region | 2 | 13 | <0.01 |

3. Fields, formats and indexing for tables to import

Data Specification: <https://cprd.com/sites/default/files/CPRD%20Aurum%20Data%20Specification%20v2.4.pdf>

NB: 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 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. **Haven’t worried about this in MySQL.**

a) Data tables

i) Drug issue table

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Field name full** | **Field name in files** | **CPRD format (from Data Specification)** | | **MySQL format** | **NULLs ?** | **Indexing** |
| 1 | Patient identifier | patid | TEXT | 6-19 numeric characters | BIGINT UNSIGNED |  | Index |
| 2 | Issue record identifier | issueid | TEXT | Up to 19 numeric characters | BIGINT UNSIGNED |  | Primary key |
| 3 | CPRD Practice identifier | pracid | INTEGER | 5 | MEDIUMINT |  |  |
| 4 | Problem observation identifier | probobsid | TEXT | Up to 19 numeric characters | BIGINT UNSIGNED | Yes |  |
| 5 | Drug record identifier | drugrecid | TEXT | Up to 19 numeric characters | BIGINT UNSIGNED | Yes |  |
| 6 | Event date | issuedate | DATE | DD/MM/YYYY | DATE (will need to convert) |  | Index |
| 7 | Entered date | enterdate | DATE | DD/MM/YYYY | DATE (will need to convert) |  |  |
| 8 | Staff identifier | staffid | TEXT | Up to 10 numeric characters | BIGINT | Yes |  |
| 9 | Drug code identifier | prodcodeid | TEXT | 6-18 numeric characters | BIGINT |  | Index |
| 10 | Dosage identifier | dosageid | TEXT | 64 characters | CHAR(64) | \* |  |
| 11 | Quantity | quantity | DECIMAL | 9.3 (9 digits, 3 after decimal place) | DECIMAL(9,3) |  |  |
| 12 | Quantity unit identifier | quantunitid | INTEGER | 2 | TINYINT | Yes |  |
| 13 | Course duration in days | duration | INTEGER | 10 | BIGINT |  |  |
| 14 | Estimated NHS cost | estnhscost | DECIMAL | 10.4 (10 digits, 4 after decimal place) | DECIMAL(10,4) |  |  |

\*None have this missing, but from looking at a subset of June 2020 download, approximately 50% have a dosageid which isn’t in the common\_dosages lookup

ii) Observation table

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Field name full** | **Field name in files** | **CPRD format (from Data Specification)** | | **MySQL format** | **NULLs ?** | **Indexing** |
| 1 | Patient identifier | patid | TEXT | 6-19 numeric characters | BIGINT UNSIGNED |  | Index |
| 2 | Consultation identifier | consid | TEXT | Up to 19 numeric characters | BIGINT UNSIGNED | Yes |  |
| 3 | CPRD Practice identifier | pracid | INTEGER | 5 | MEDIUMINT |  |  |
| 4 | Observation identifier | obsid\* | TEXT | Up to 19 numeric characters | BIGINT UNSIGNED |  | Primary key\* |
| 5 | Event date | obsdate | DATE | DD/MM/YYYY | DATE (will need to convert) | Yes | Index |
| 6 | Entered date | enterdate | DATE | DD/MM/YYYY | DATE (will need to convert) |  |  |
| 7 | Staff identifier | staffid | TEXT | Up to 10 numeric characters | BIGINT | Yes |  |
| 8 | Parent observation identifier | parentobsid | TEXT | Up to 19 numeric characters | BIGINT UNSIGNED | Yes |  |
| 9 | Medical code | medcodeid | TEXT | 6-18 numeric characters | BIGINT | Yes | Index |
| 10 | Value | **testvalue:** changed from value as MySQL already has value | NUMERIC | 19.3 | DECIMAL(19,3) | Yes |  |
| 11 | Numeric unit identifier | numunitid | INTEGER | 10 | BIGINT | Yes |  |
| 12 | Observation type identifier | obstypeid | INTEGER | 5 | MEDIUMINT |  |  |
| 13 | Numeric range low | numrangelow | NUMERIC | 19.3 | DECIMAL(19,3) | Yes |  |
| 14 | Numeric range high | numrangehigh | NUMERIC | 19.3 | DECIMAL(19,3) | Yes |  |
| 15 | Problem observation identifier | probobsid | TEXT | Up to 19 numeric characters | BIGINT UNSIGNED | Yes |  |

\*Discovered during test on Pearson that there is a duplicate row (all fields duplicated) with obsid 8544593954 in males data

iii) Patient table

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Field name full** | **Field name in files** | **CPRD format (from Data Specification)** | | **MySQL format** | **NULLs ?** | **Indexing** |
| 1 | Patient identifier | patid | TEXT | 6-19 numeric characters | BIGINT UNSIGNED |  | Primary key |
| 2 | CPRD practice identifier | pracid | INTEGER | 5 | MEDIUMINT |  |  |
| 3 | Usual GP | usualgpstaffid | TEXT | Up to 10 numeric characters | BIGINT | Yes |  |
| 4 | Gender | gender | INTEGER | 3 | SMALLINT |  |  |
| 5 | Year of birth | yob | INTEGER | 4 | SMALLINT |  |  |
| 6 | Month of birth | mob | INTEGER | 2 | TINYINT | Yes |  |
| 7 | Date of death | emis\_ddate | DATE | DD/MM/YYYY | DATE (will need to convert) | Yes |  |
| 8 | Registration start date | regstartdate | DATE | DD/MM/YYYY | DATE (will need to convert) |  |  |
| 9 | Patient type | patienttypeid | INTEGER | 5 | MEDIUMINT |  |  |
| 10 | Registration end date | regenddate | DATE | DD/MM/YYYY | DATE (will need to convert) | Yes |  |
| 11 | Acceptable flag | acceptable | INTEGER | 1 | BOOL |  |  |
| 12 | CPRD death date | cprd\_ddate | DATE | DD/MM/YYYY | DATE (will need to convert) | Yes | Index |

Notes from looking at full table on MySQL on 11/01/2021:

* Everyone with emis\_ddate also has cprd\_ddate
* 1629 people with no emis\_ddate have cprd\_ddate – use this
* 5 people with cprd\_ddate do not have regenddate – so do need to always check death

iii) Practice table

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Field name full** | **Field name in files** | **CPRD format (from Data Specification)** | | **MySQL format** | **NULLs ?** | **Indexing** |
| 1 | CPRD practice identifier | pracid | INTEGER | 5 | MEDIUMINT |  | Primary key |
| 2 | Last Collection Date | lcd | DATE | DD/MM/YYYY | DATE (will need to convert) |  |  |
| 3 | Up-to-standard date: **not currently populated** | uts | DATE | DD/MM/YYYY | DATE (will need to convert) | Yes |  |
| 4 | Region | region | INTEGER | 5 | MEDIUMINT | Yes |  |

b) Lookup tables

i) Medical Dictionary

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Field name** | **CPRD format (from Data Specification)** | | **MySQL format** | **NULLs ?** | **Indexing** |
| 1 | medcodeid | TEXT | 6-18 numeric characters | BIGINT |  | Primary key |
| 2 | term | TEXT | Up to 255 characters | VARCHAR(255) |  | Index |
| 3 | originalreadcode\* | TEXT | Up to 100 characters | VARCHAR(100) |  | Index |
| 4 | cleansedreadcode\* | TEXT | Up to 7 characters | VARCHAR(7) | Yes |  |
| 5 | snomedctconceptid | TEXT | Up to 19 numeric characters | BIGINT UNSIGNED |  |  |
| 6 | snomedctdescriptionid | TEXT | Up to 19 numeric characters | BIGINT UNSIGNED |  |  |
| 7 | releaseid: **not currently populated. Changed name from release as already release variable in MySQL** | TEXT | Up to 100 characters | VARCHAR(100) | Yes |  |
| 8 | emiscodecategoryid | INTEGER | 2 | TINYINT |  |  |

\*Will need to be case sensitive

ii) Product Dictionary

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Field name** | **CPRD format (from Data Specification)** | | **MySQL format** | **NULLs ?** | **Indexing** |
| 1 | prodcodeid | TEXT | 6-18 numeric characters | BIGINT |  | Primary key |
| 2 | dmdid | TEXT | Up to 19 numeric characters | BIGINT UNSIGNED | Yes |  |
| 3 | termfromemis | TEXT | Up to 255 characters | VARCHAR(255) |  |  |
| 4 | productname\* | TEXT | Up to 999 characters | VARCHAR(999) | Yes | Index |
| 5 | formulation | TEXT | Up to 999 characters | VARCHAR(999) | Yes |  |
| 6 | routeofadministration | TEXT | Up to 999 characters | VARCHAR(999) | Yes |  |
| 7 | drugsubstancename | TEXT | Up to 999 characters | VARCHAR(999) | Yes | Index |
| 8 | substancestrength | TEXT | Up to 999 characters | VARCHAR(999) | Yes |  |
| 9 | bnfchapter | TEXT | Up to 999 characters | VARCHAR(999) | Yes | Index |
| 10 | releaseid: **not currently populated. Changed name from release as already release variable in MySQL** | TEXT | Up to 100 characters | VARCHAR(100) | Yes |  |

\* On import: replace productname with termfromemis if productname missing. Most termfromemis = productname + name of manufacturer

iii) common\_dosages (data formats not in Data Specification; extracted from May 2020 lookups)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Field name** | **CPRD format** | **MySQL format** | **NULLs ?** | **Indexing** |
| 1 | dosageid | 64 characters | CHAR(64) not case sensitive |  | Primary key |
| 2 | dosage\_text | Up to 999 characters | VARCHAR(999)\* |  |  |
| 3 | daily\_dose | Decimal 15,9 | DECIMAL(15,9) |  |  |
| 4 | dose\_number | Decimal 14,7 | DECIMAL(14,7)\*\* |  |  |
| 5 | dose\_unit | Up to 7 present | VARCHAR(20) | Yes |  |
| 6 | dose\_frequency | Decimal 9,6 | DECIMAL(9,6) |  |  |
| 7 | dose\_interval | Decimal 11,8 | DECIMAL(11,8) |  |  |
| 8 | choice\_of\_dose | Integer 1 | TINYINT |  |  |
| 9 | dose\_max\_average | Integer 1 | TINYINT |  |  |
| 10 | change\_dose | Integer 1 | TINYINT |  |  |
| 11 | dose\_duration | Decimal 4,1 | DECIMAL(4,1) |  |  |

\*When tested putting this table on MySQL in January 2021, realised that MySQL uses backslash (‘\’) as default escape character. Some dosage\_text here contain backslashes so need to add ‘ESCAPED BY '\b'’ when importing (between field terminators and line terminators)

\*\*Some dose numbers have been converted to exponents e.g. 8.765432E+07. None of these are true doses – replace with NULL during import to avoid errors with fitting the above specified MySQL format

iv) EmisCodeCat (not in Data Specification; extracted from May 2020 lookups)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Field name** | **CPRD format** | **MySQL format** | **NULLs ?** | **Indexing** |
| 1 | emiscodecatid | 2 digit integer | TINYINT |  | Primary key |
| 2 | description | Up to 43 characters | VARCHAR(100) |  |  |

v) Gender (not in Data Specification; extracted from May 2020 lookups)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Field name** | **CPRD format** | **MySQL format** | **NULLs ?** | **Indexing** |
| 1 | genderid | 1 digit integer | TINYINT |  | Primary key |
| 2 | description | 1 character | VARCHAR(1) |  |  |

vi) NumUnit (not in Data Specification; extracted from May 2020 lookups)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Field name** | **CPRD format** | **MySQL format** | **NULLs ?** | **Indexing** |
| 1 | numunitid | 5 digit integer | MEDIUMINT |  | Primary key |
| 2 | description | Up to 50 characters | VARCHAR(100)\* |  |  |

\*When tested putting this table on MySQL in January 2021, realised that MySQL uses backslash (‘\’) as default escape character. Descriptions here contain backslashes so need to add ‘ESCAPED BY '\b'’ when importing (between field terminators and line terminators)

vii) ObsType (not in Data Specification; extracted from May 2020 lookups)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Field name** | **CPRD format** | **MySQL format** | **NULLs ?** | **Indexing** |
| 1 | obstypeid | 2 digit integer | TINYINT |  | Primary key |
| 2 | description | Up to 15 characters | VARCHAR(100) |  |  |

viii) PatientType (not in Data Specification; extracted from May 2020 lookups)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Field name** | **CPRD format** | **MySQL format** | **NULLs ?** | **Indexing** |
| 1 | patienttypeid | 2 digit integer | TINYINT |  | Primary key |
| 2 | description | Up to 25 characters | VARCHAR(100) |  |  |

ix) QuantUnit (not in Data Specification; extracted from May 2020 lookups)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Field name** | **CPRD format** | **MySQL format** | **NULLs ?** | **Indexing** |
| 1 | quantunitid | 2 digit integer but nearly 100 | SMALLINT |  | Primary key |
| 2 | description | Up to 31 characters | VARCHAR(100) |  |  |

x) Region (not in Data Specification; extracted from May 2020 lookups)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Field name** | **CPRD format** | **MySQL format** | **NULLs ?** | **Indexing** |
| 1 | regionid | 2 digit integer | TINYINT |  | Primary key |
| 2 | description | Up to 24 characters | VARCHAR(100) |  |  |

4. Other notes

* Don’t import first line of table (IGNORE 1 LINES)
* If field has empty values, replace with NULL at import – have tested with HbA1c codes – look at this code for how to convert dates/deal with potentially NULL dates
* If make new lookup tables e.g. of HbA1c medcodes, need to remember to add primary key/other indexes
* In Rob Challen’s code, he drops table if exists before creating each one
* Rob Challen says 'We have not specified the foreign keys as there is no evidence that they improve SELECT performance'
* Use MyISAM engine for raw data tables?
* Decided to use case sensitive schema, so names of objects (tables + fields within them) are case sensitive
* Collation for text fields: only Read codes need to be case sensitive, rest do not
* Set up in a way which allows removal of 4 x main data tables afterwards (fine to leave derived tables on there?)
* Need to think about index combinations...not sure if what I’ve suggested above is a good idea
  + Will definitely want to index medcodes/prodcodes in Observations/Drug Issue tables
* Logging on MySQL in case crashes/errors?

For Observations and Drug Issues tables – so large so took a very long time to import into MySQL:

* Using command line:
  + Copied files without header row (awk -F'\t' 'NR>1' oldfile.txt > newfile.txt)
  + Sorted by primary key (obsid / issueid) (sort –k2n / -k4n oldfile.txt > newfile.txt)
  + Split…..
* MySQL:
  + Imported one by one
  + Switched off foreign key checks (SET foreign\_key\_checks = 0)