# UK Biobank

## Primary Care Data for COVID-19 Research (deprecated resource)

---

Version 3.1

www.ukbiobank.ac.uk
March 2024

This document version has been updated to indicate that the document is deprecated as the primary care data for COVID-19 research are no longer available.
**Researchers who previously accessed the data may only use it for COVID-19 related research.**

# Contents

# 1 Overview of this release

## 1.1 Data available for COVID-19 research

This initial release of primary care data for COVID-19 research contained data for approximately 409,000 participants in England covering the primary care (GP) practices with TPP (https://www.tpp-uk.com/) or EMIS (https://www.emishealth.com/) as their data system supplier. Primary care data from Scotland and Wales were not published.

**The primary care data for COVID-19 related research are no longer available for researchers to request as the COPI notice covering the data has expired. Researchers who accessed the data prior to its withdrawal may use the data for COVID-19 research ONLY. See section 3 for further details.**

This resource contains details of the dataset for reference purposes only, for researchers who accessed it prior to its removal.

An overview of all of the linked health data available from UK Biobank for COVID-19 research can be found on the Essential Information section of Showcase.

## 1.2 Downloading primary care data for COVID-19 research purposes

The COVID-19 primary care data tables were:

- covid19_emis_gp_clinical (clinical data from GP practices with EMIS data system supplier)
- covid19_emis_gp_scripts (drug data from GP practices with EMIS data system supplier)
- covid19_tpp_gp_clinical (clinical data from GP practices with TPP data system supplier)
- covid19_tpp_gp_scripts (drug data from GP practices with TPP data system supplier)

Section 2 describes primary care data in general as well as UK Biobank's previous primary care data release. Section 3 describes the permitted uses of this data release. Section 4 gives further information on the structure of the available data and the coding systems used, and Section 5 gives details of the (minimal) data cleaning undertaken by UK Biobank. Section 6 gives further details on the Data Portal.

# 2 Primary care data

## 2.1 Primary care in the UK

Within the UK healthcare setting, individuals seeking advice or treatment for a health concern normally first meet with a family physician (known as a General Practitioner, or GP) or a nurse (for example, a Nurse Practitioner) at their local general practice. GPs can refer patients who require more specialised treatment, or further tests, to hospital or another community-based service.

There is a wealth of information available within primary care records. Some illnesses are managed entirely within a primary care setting, and, most secondary care interactions are reported back to the GP and entered into their electronic medical record.

The term 'primary care' is sometimes used more broadly to include other healthcare professionals such as pharmacists, dentists and opticians. The UK Biobank primary care data relates only to data recorded by health care professionals working at general practices.

## 2.2 The nature of primary care data

Primary care data are 'real world' administrative data. Such routinely collected administrative data have enormous potential to support research with far reaching benefits to human health.

By their nature, analysing and interpreting these data within the context of health research requires careful consideration of their content, structure and crucially, understanding that they were collected for an entirely different purpose: recording the delivery of patient care in thousands of different centres across the UK operating within their own NHS systems.

## 2.3 UK Biobank previously released primary care data

UK Biobank has been liaising with system suppliers and other intermediaries to obtain primary care data for UK Biobank participants, all of whom have provided written consent for linkage to their health-related records.

Primary care data from multiple providers encompassing approximately 45% of the UK Biobank cohort - i.e. approximately 230,000 participants - were made available in September 2019 and are accessible via the Data Showcase. Further details of these data are found in Resource 591 in Category 3000 on Showcase. They are not restricted for COVID-19 research purposes and can be used for all approved research.

# 3 Using the primary care data for COVID-19 research

## 3.1 Basis for access and permitted uses of data

In March 2020, the Secretary of State for Health and Social Care issued a notice under the Control of Patient Information (COPI) Regulations to all GP practices in England (using the TPP or EMIS systems), to instruct them to release the relevant primary care data to UK Biobank for purposes related to the outbreak of COVID-19.

Purposes related to the outbreak of COVID-19 include, but are not limited to, the following:

- Understanding COVID-19 and risks to public health, trends in COVID-19 and such risks, and controlling and preventing the spread of COVID-19 and such risks;
- Identifying and understanding information about patients or potential patients with or at risk of COVID-19, information about incidents of patient exposure to COVID-19 and the management of patients with or at risk of COVID-19 including: locating, contacting, screening, flagging and monitoring such patients and collecting information about and providing services in relation to testing, diagnosis, self-isolation, fitness to work, treatment, medical and social interventions and recovery from COVID-19;
- Understanding information about patient access to health services and adult social care services and the need for wider care of patients and vulnerable groups as a direct or indirect result of COVID-19 and the availability and capacity of those services or that care;
- Monitoring and managing the response to COVID-19 by health and social care bodies and the Government, including providing information to the public about COVID-19 and its effectiveness and information about capacity, medicines, equipment, supplies, services and the workforce within the health services and adult social care services;
- Delivering services and information to patients, clinicians, the health services and adult social care services workforce and the public about and in connection with COVID-19, including the provision of information, fit notes and the provision of health care and adult social care services;
- Research and planning in relation to COVID-19.

The extracts of primary care data that UK Biobank received under the COPI regulations were very similar to those received previously as part of UK Biobank's ongoing record linkage programme (i.e. they contain data on coded diagnoses, symptoms, medications, referrals etc.). The extracts were not restricted to participants with suspected or confirmed COVID-19.

**Aside from containing more recent records, the key difference to this extract is that these primary care data are only to be used for COVID-19 related research. The COPI notice has now expired and so these data may no longer be released.**

## 3.2 Researcher obligations

We would like to take this opportunity to remind researchers that all research outputs (including pre-prints/publications and other results posted on social media) should be sent to UK Biobank prior to public release. This is not to obtain UK Biobank approval (as this is not required), but so that we are fully aware of any article that may generate media interest.

Researchers should also ensure that the communications teams within their own institutes are aware of the work and that they notify the UK Biobank communications team of any press activity.

# 4 Format of the COVID-19 primary care data

The COVID-19 primary care tables made available in this release contain data on:

- coded clinical events – including diagnoses, history, symptoms, lab results, and procedures;
- prescriptions issued by GPs;
- a range of administrative codes (e.g. referrals to specialist hospital clinics).

Non-coded, unstructured data such as free-text entries, referral letters, etc. are not included.

Information on participant registrations, drug names and the quantity of medication issued is not available in this initial data release, although we intend to make this available in due course.

Data from deceased patients are included, although please bear in mind that no data cleaning or detailed analysis has been undertaken comparing dates of clinical or prescription events with that of date of death.

Researchers should also be aware that an absence of primary care records for a period of time may reflect the participant being registered at a practice using a different software system, rather than a period with no primary care consultations, and that the completeness of data transfer when a patient moves between practices (and system suppliers) is unknown.

In addition, researchers should be aware that prescriptions issued by GPs cannot be assumed to have been dispensed by a pharmacy in all cases.

## 4.1 System suppliers and their coding classifications

Table 1 below summarises the coding classification systems for clinical event and prescriptions that are used by TPP and EMIS, the two main computer system suppliers in England. TPP provides the SystmOne practice management system, and EMIS Health provides the EMIS Web practice management system.

**Table 1. System suppliers and their coding classifications**

| | EMIS GP system supplier | TPP GP system supplier |
|---|---|---|
| **Clinical** | SNOMED CT<br>Local EMIS codes<br>Local EMIS test request codes | Clinical Terms Version 3 (CTV3, Read v3)<br>Local TPP codes |
| **Prescription** | Dictionary of medicines and devices (dm+d)<br>Local EMIS codes | dm+d |

SNOMED CT and dm+d look-up tables should be downloaded from the UKTC Terminology Reference data Update Distribution (TRUD) service (see Section 4.3.1), as we cannot provide these directly. Read v3 codes are available for download via the Data Showcase. Note that Read v3 codes are no longer being updated via TRUD and have now been replaced with SNOMED CT; see Section 4.3.2.

Each of the coding classification systems used in this release are described below with links to resources for more information. Please note that primary care coding systems are dynamic and subject to a range of potential biases and fluctuations over time due to national and local policy initiatives and local processes and procedures. Their completeness and accuracy, relative to the actual health experiences of the individuals represented in the coded data, cannot be assumed and is expected to differ between systems and over time. The accuracy of code lists, definitions and maps should be verified by specialists as part of any analysis undertaken on these data.

**Table 2. Encodings used to classify clinical coding systems used in GP data**

| Encoding | Coding classification system |
|---|---|
| 0 | CTV3 code |
| 1 | Local TPP code |
| 2 | SNOMED CT |
| 3 | Local EMIS code |
| 5 | Local EMIS test request code (OLTR) |
| 6 | dm+d |

## 4.2 Coding systems used in EMIS

### 4.2.1 SNOMED CT

SNOMED CT ('Systemized Nomenclature of Medicine Clinical Terms') is a structured vocabulary of clinical phrases and terms ('*concepts*') that aims to characterise all healthcare processes, including diagnosis and procedures, symptoms, family history, allergies, assessment tools, observations, medications, and devices.

Updates to clinical codes are released every six months via TRUD, although more frequent micro-releases have been occurring in response to the COVID-19 pandemic. The 6- to 18-digit numerical codes are usually issued in the arbitrary order in which components are added to SNOMED CT over time, so within a clinical subcategory there is often heterogeneity in the value of codes rather than compliance to a specific identifiable range.

Instructions on downloading SNOMED CT releases from TRUD and identifying the relevant files are provided in Appendix A. The NHS Digital SNOMED CT Browser can be used to identify specific concept codes ('*SCTID*') of interest. For more in-depth knowledge NHS Digital provide a comprehensive online presentation on SNOMED CT release files.

The latest information on COVID-19 terminology and future releases are available from the NHS Digital Delen Platform. As of August 2020, there were four refsets of COVID-19 related codes available for download from TRUD, on: presenting complaints, health issues, record extraction, and, procedures.

### 4.2.2 EMIS local codes

EMIS data also contain bespoke code lists. Where possible, the system providers have mapped these onto SNOMED CT codes. Nevertheless, some local codes remain. These local codes pertain to clinical events, online test requests (OLTRs), test results and prescriptions and appear in both the clinical and prescription tables. A list of EMIS local codes that are present in the current extract and their definitions can be found in Data-Coding 7689 (clinical codes) and Data-Coding 7678 (prescription codes). Some EMIS-supplied codes (i.e. labelled as Read v2 codes) were not able to be verified against known coding systems and were therefore redacted for this release.

### 4.2.3 dm+d (Dictionary of Medicines and Devices)

The EMIS prescription data are coded using either EMIS local codes (see above) or dm+d codes, which was developed for use throughout the NHS to identify specific medicines and devices used in the treatment of patients and consists of a dictionary containing unique identifiers and associated text descriptions. The dm+d dictionary is available here.

The dm+d model consists of five components:

- A Virtual Therapeutic Moiety (VTM) – the substances intended for use in the treatment of a patient;
- Virtual Medicinal Product (VMP) – an abstract concept representing the properties of one or more clinically equivalent Actual Medicinal Products (AMPs);
- Actual Medicinal Product (AMP) – a single dose unit of an actual product known to have been available from a specific supplier;
- Virtual Medicinal Product Pack (VMPP) – an abstract concept representing the properties of one or more quantitatively equivalent Actual Medicinal Product Packs (AMPPs);
- Actual Medicinal Product Pack (AMPP) – the packaged product supplied for direct patient use.

An example of the dm+d component structure for a packet containing 56 tablets of Yaltormin 500mg is shown below in Table 4. Note the generic name appears in the VTM, VMP and VMPP dm+d components while the brand name is used in the AMP and AMPP components.

**Table 3. Example of dm+d codes, components and descriptions**

| dm+d code | dm+d | Description |
|---|---|---|
| 109081006 | VTM | Metformin |
| 386047000 | VMP | Metformin 500mg modified-release tablets |
| 355475110000011 | AMP | Yaltormin SR 500mg tablets (Wockhardt UK Ltd) |
| 899061100000110 | VMPP | Metformin 500mg modified-release tablets 56 tablets |
| 355479110000011 | AMPP | Yaltormin SR 500mg tablets (Wockhardt UK Ltd) 56 |

More information about the dm+d model is available as part of a series of short webinars available from the TRUD website and code lookups are also available (registration required). The dm+d codes are provided in XML files. To convert them into tabular format, the dm+d XML Transformation Tool (downloadable from TRUD) is required. Details on using the dm+d XML Transformation Tool (Windows only) can be found in Appendix B. The NHS Business Services Authority (NHSBSA) also provides a web-based dm+d browser.

As dm+d codes are updated on a regular basis, we have not provided encodings on Data Showcase for these codes and instead refer researchers to the NHSBSA dm+d Browser and corresponding lookup files on TRUD.

## 4.3 Coding systems used in TPP

### 4.3.1 CTV3

Read codes are a coded thesaurus of clinical terms used in primary care since 1985. There are two versions: version 2 (Read v2) and version 3 (CTV3 or Read v3). Both provide a standard vocabulary for clinicians to record patient findings and procedures.

TPP clinical data are coded using CTV3 together with some additional codes local to TPP (see below). The final update of CTV3 was in April 2018 and the system is now no longer in active use (nor is the Read Browser), owing to the phased introduction of SNOMED CT. At the time of writing, CTV3 definition files, together with a UK Read code browser, are available for download via the NHS Digital Technology Reference Data Update Distribution (TRUD), however they are scheduled for removal in 2020. As well as the TRUD resources, CTV3 codes are available via the Data Showcase using Data-Coding 7128.

### 4.3.2 TPP local codes

TPP data also contain bespoke code lists. Where possible, the system providers have mapped these onto more widely known coding schema (i.e. CTV3), although some local codes remain. Since CTV3 has

not been updated since April 2018, new codes - including COVID-19 specific codes related to its diagnosis, testing, symptoms, referrals, categories of risk for self-isolation etc. - are captured by 'local' TPP-specific codes.

Researchers are strongly advised to investigate these local code lists (for example to identify whether there is temporal and/or geographical variation in their usage, and to assess the possibility of duplication with CTV3) and interpret their findings accordingly.

### 4.3.3 dm+d (Dictionary of Medicines and Devices)

Please see information in the description of dm+d codes provided for EMIS data above. Prescription codes in the TPP data consist of AMP and VMP codes only.

## 4.4 Structure of primary care tables

The clinical and prescription data are provided in two separate tables on the Data Portal. There is no specific key field linking entries between the clinical and prescription data. Linking records between the two tables will require matching participant identifiers and dates.

**Table 4. Structure of clinical tables**

| Table Name | Field Name | Description | Encoding |
|---|---|---|---|
| covid19_emis_gp_clinical | eid | Participant identifier | NA |
| | event_dt | Date clinical code entered | Special codes in Data-Coding 819 |
| | code | Clinical code | SNOMED codes available from TRUD<br>EMIS Local Clinical Code List Data-Coding 7689 |
| | code_type | SNOMED or EMIS Local code | Data-Coding 3175 |
| | value | Value recorded | Special codes in Data-Coding 2360 |
| | unit | Unit recorded for value | Special codes in Data-Coding 1176 |
| covid19_tpp_gp_clinical | eid | Participant identifier | NA |
| | event_dt | Date clinical code entered | Special codes in Data-Coding 819 |
| | code | Clinical code | CTV3: Data-Coding 7128<br>Local TPP: Data-Coding 8708 |
| | code_type | CTV3 or local TPP code | Data-Coding 3175 |
| | value | Value recorded | Special codes in Data-Coding 5702 |

The clinical code table contains data on primary care events, such as consultations, diagnoses, history, symptoms, procedures, laboratory tests and administrative information. While some data are included here on immunisations, this information is not comprehensive (as most immunisation data are provided separately and are not included in this release). Where available, a value field is provided that may give further details. For TPP data, no unit information was available to contextualise the value field.

**Table 5. Structure of prescription tables**

| Table Name | Field Name | Description | Encoding |
|---|---|---|---|
| covid19_emis_gp_scripts | eid | Participant identifier | NA |
| | issue_date | Date clinical code entered | Special codes in Data-Coding 819 |
| | code | Clinical code | dm+d codes available from TRUD<br>EMIS Local Prescription Code List: Data-Coding 7678 |
| | code_type | dm+d or EMIS Local code | Data-Coding 3175 |
| covid19_tpp_gp_scripts | eid | Participant identifier | NA |
| | issue_date | Date clinical code entered | Special codes in Data-Coding 819 |
| | dmd_code | Clinical code | dm+d codes available from TRUD<br>Special codes in Data-Coding 4214 |

The prescription table contains information on the issuing of prescription medication. Researchers should be aware that prescriptions issued by GPs cannot be assumed to have been dispensed by a pharmacy in all cases.

# 5 Data quality & cleaning

## 5.1 Data Quality

We are making this release of this COVID-19 primary care data available to researchers in a form as close as possible to that provided to UK Biobank by the system supplier. This is to avoid introducing a potential systematic error or bias through data cleaning efforts. Please be aware therefore that this is not a complete and error-free dataset, which may have implications for the conclusions you draw when interpreting your research.

As previously noted, this COVID-19 primary care release only includes data from a subset of participants, and it should not be assumed that analyses conducted on this subset can be generalised to the entire UK Biobank cohort, or the UK population as a whole. Appropriate analytical techniques must be employed to account for missing or unreliable data. This includes identifying:

- erroneous information, such as dates or codes entered incorrectly
- inconsistent information, such as variation in timing or content of records between sources
- data gaps, absent data, or variation in completeness of available data

To prevent against possible participant identification, we have removed potentially disclosive or sensitive codes and values from the data, as outlined below.

## 5.2 Data Cleaning

### 5.2.1 Redacting potentially sensitive codes

Some clinical codes have the potential to be sensitive and/or identifying. For example, rare occupation codes in combination with other information held by UK Biobank about a participant may potentially disclose a participant's identity. As such, we have redacted rare occupation codes (defined as those that appeared in no more than five records). Code descriptions have also been reviewed using keyword searches to identify and redact other codes that could be identifying, and/or are otherwise deemed to be sensitive and/or not relevant to Covid-19 research (see Appendix C for more details and a list of subcategories of redacted codes.)

### 5.2.2 Missing prescription codes

Any drugs prescribed using electronic prescribing in the UK are linked to dm+d, which means most drugs are coded using this schema in the dataset. Nevertheless, in the TPP data, the dm+d code may be absent for some food-related items, bandages, older or obsolete drugs, and some rarer formulations, or when the item is mapped to multiple dm+d codes. In these circumstances, the data has been recoded using Data-Coding 4214.

While EMIS data generally provided local EMIS codes in the case of missing dm+d codes, some codes are missing and have been recoded. EMIS prescription data also contains some unverifiable Read v2 codes; please see Section 5.2.6 Unverifiable codes. EMIS local prescription codes as well as recodes are found in Data-Coding 7678.

### 5.2.3 Numerical free-text fields

Manual checks on the contents of the 'value' field in the clinical events table have been carried out to check for potentially disclosive values, such as a participant's date of birth or phone number. These values were recoded, as well as values that accompanied codes that were deemed potentially sensitive, identifying or relating to someone other than the participant. These recodes are found in Data-Coding 5702 for TPP and 2360 for EMIS.

### 5.2.4 Dates

We have altered dates in relation to participants' date of birth, as follows:

- Where dates were missing they have been recoded to 01/01/1900;
- Where a clinical event or prescription date occurs before the participant's date of birth, it has been altered to 01/01/1901 (note that this also recodes dates such as 01/01/1900 from the raw data, which may have been intended to mean "unknown date");
- Where the date matches the participant's date of birth, it has been altered to 02/02/1902;
- Where the date occurs after the participant's date of birth but is in the year of their birth, it has been altered to 03/03/1903;

- Where a future date has been entered, it has been altered to 07/07/2037 (as these are likely to have been entered as a place-holder or other system default).

These re-codes are found in Data-Coding 819.

### 5.2.5 Duplicate records

Records that are exact duplicates of another record (i.e. all released fields are exactly the same) have been removed from the data.

### 5.2.6 Unverifiable codes

In the clinical data provided by EMIS we have been unable to verify a small subset of translated SNOMED CT codes because they are not listed as clinical concept codes in the SNOMED CT International Edition or UK Extension releases. Where available, we have released the corresponding EMIS local codes for these invalid SNOMED codes. Where these were unavailable, the unverifiable SNOMED code was redacted. In addition, a small proportion of clinical codes (<1%) in the EMIS GP data were listed as Read v2 codes, but do not exist on published Read v2 code lists from NHS Digital. As such, these codes have been redacted from this data release.

### 5.2.7 Other notes on codes

Note that CTV3, dm+d and SNOMED CT use hierarchical structures that are not expressed in the codes themselves. To ensure you have all the relevant codes for a condition or medication, we recommend you use the NHS Read Browser, dm+d Browser or SNOMED CT Browser (see Section 4.3 for more information). A hierarchy guide for local TPP codes or EMIS codes was not available at the time of writing. In particular, care should be taken with CTV3, since nested within the hierarchy there are sometimes specific codes that indicate a person did not have the condition.

† Please be aware that there are multiple clinical codes for COVID-19 as confirmed by laboratory test, or by diagnostic criteria, as well as other codes in the data that pertain to suspected COVID-19, exposure to others with COVID-19, test requests for COVID-19, and more. We would therefore encourage use of these multiple codes rather than restrict selection to the above examples alone.

# Appendix A. Downloading SNOMED CT code lists

SNOMED CT International and UK Clinical Extension code lists are downloaded as tab delimited text files. In order to get the full UK Edition of SNOMED CT codes researchers must append the Concepts file from the International release to the Concepts file of the UK Clinical Extension. Both International and UK files are included in a single release.

The steps to downloading the files are as follows:

1. Register and sign in to NHS Digital TRUD;
2. Navigate to SNOMED CT UK Edition;
3. Select 'UK SNOMED CT Clinical Edition, RF2: Full, Snapshot & Delta' and subscribe to the licenses for this item;
4. Select 'Download releases' and then 'Download' for the latest pack. Releases are listed in reverse date order. For example:
   File name "SNOMEDCT2_30.0.0_20200805000001" means:
   - SNOMED CT R2 file (NB. R2 files were released in October 2013 and include all components from R1 released in April 2004);
   - 30th version of the release;
   - Released on 5th August 2020 at 1 second passed midnight.
5. Use corresponding files from both the 'SnomedCT_InternationalRF2_PRODUCTION…' and 'SNOMEDCT_UKClinicalRF2_PRODUCTION…' subfolders. Each subfolder includes an identical filing structure.

## Navigating the main subfolders within a single SNOMED CT release

We recommend using the Full release files when downloading SNOMED CT codes for the first time. Definitions for the three different file types are as follows:

| | |
|---|---|
| ***Full*** | include a full history of every version of every component ever released. If there have been changes to a component a new row will be added, and a new status. Full data is recorded from 1st April 2004. |
| ***Snapshot*** | contain the current state of every component ever released, i.e. 1 row per component. |
| ***Delta*** | include all new components that have been added or changed, or inactivated, i.e. the difference between this full release and the previous full release |

## Identifying the main data files

Within each of the Full, Snapshot and Delta release folders are 'Terminology' and 'Refset' subfolders. The Terminology folder holds the main data files, in particular Concepts and Descriptions respectively:

- The Concept .txt file includes all SNOMED CT concept codes (field name: 'id') within the International release ('sct2_Concept_Full_INT_*YYYYMMDD*.txt') and UK Clinical Extension respectively ('sct2_Concept_Full_GB1000000_*YYYYMMDD*.txt'). As stated above both files need appending to one another in order to achieve the full UK Edition of SNOMED CT.

- The Description .txt files include the concept codes ('conceptId') alongside the text description ('term') for each concept, and will therefore also be useful for researchers.

The Refsets subfolder contains .txt files that list SNOMED concepts relating to specific clinical areas. However, these can be separately downloaded from TRUD as .xlsx files that include both the concept names and descriptions in a more accessible format (i.e. navigate to 'SNOMED CT subset members in spreadsheet view'.)

For more information see NHS Digital's online presentation on 'An Introduction to the SNOMED CT Release Files', and the International Health Terminology Standards Development Organisation's SNOMED CT Technical Implementation Guide.

# Appendix B. Using the dm+d XML Transformation Tool

To extract dm+d codes in tabular format, both the latest release of dm+d XML file and the dm+d XML Transformation Tool must be downloaded from TRUD. Please note that these files must be registered for and downloaded separately. More information on downloading and understanding dm+d files is available as part of a series of short webinars from TRUD (registration required).

Please note that the dm+d XML Transformation Tool is not a UK Biobank resource and we are unable to provide guidance or give assistance with trouble-shooting.

Both files are downloaded in .zip format. To use the transformation tool to create tabular code lookups, follow the steps below:

1. Extract the transformation tool (uk_dmdextract_*n.n.n_yyyymmdd*000001.zip) to the C:\ drive. This will create the folder *dmd_extract_tool* in the C:\ drive. Move the zipped file with the dm+d content (nhsbsa_dmd_*n.n.n_yyyymmdd*000001.zip) into the directory C:\dmd_extract_tool\XMLToUnzipInHere. If there are any old zip files in the folder, you will need to delete them.
2. Navigate back to C:\dmd_extract_tool and double-click the batch file called xml-csv.bat. This will open the command prompt and will create the tabular files in C:\dmdDataLoader. The process will take a few minutes to run. If you have Microsoft Excel installed, Excel may open and request macros to be enabled – this is to create Excel files from the generated CSV files. If Microsoft Excel does not open or you do not have it installed, the CSV tables should still be generated.
3. Once the Transformation Tool has finished running, the command prompt should close and 47 CSV files can be found in C:\dmdDataLoader\csv. There are two files containing dm+d codes found in EMIS and TPP data:
    - The lookup file for VMPs is "f_vmp_VmpType.csv".
    - The lookup file for AMPs is "f_amp_AmpType.csv" (this file also serves as a crossmap giving the corresponding VMP for each AMP).

In the EMIS data there are an extremely small number of dm+d codes that are either listed in other files ("f_vmpp_VmppType.csv" and "f_amp_AmppType.csv"), or are redundant codes in the file "f_vmp_VmpType.csv", i.e. 'VPIDPREV'. None of these are obvious treatments for COVID-19 associated symptoms. For more information about the components of the dm+d model, including VMPs and AMPs, please see Section 4.2.

# Appendix C. Redaction of sensitive codes from primary care data for COVID-19 research

In order to maintain the confidentiality of participants in the UK Biobank study, and to limit the release of primary care (GP) data to that which is potentially relevant for Covid-19 research, certain clinical codes (and their accompanying values) have been redacted.

There are several reasons why a code might be redacted, which are non-mutually exclusive:

(1) Codes that could identify a participant, or are sufficiently rare that there is a risk of re-identification (e.g., rare occupations);
(2) Codes related to legal status, criminal behaviour, sensitive social information, information about non-participants, and clinical details that participants may not wish to be divulged from their medical records;
(3) Any other codes deemed potentially sensitive and not considered relevant to Covid-19 research.

Themes or subcategories of codes that fall into the above reasons include:

- Abuse (at risk of / history of / of or by the participant), e.g. female genital mutilation, living in refuge, physical restraint vulnerability identified, at risk of financial abuse, harm to others, battered baby suspect. NB. This does not include self-harm
- Adoption status, orphan/orphanage, fostering status
- Asylum seeker; immigrant
- Carer issues
- Clinical or personal information about a specific person other than the participant, e.g. mother of participant is hepatitis B positive, child with Down's syndrome in family
- Consent to treatment. NB. Exceptions are immunisations / vaccinations
- Criminal offence (or indicative of), i.e. illegal abortion or miscarriage, illegal tenancy, prison, trespasser, vandalism, child on remand, acting as police witness
- Dates of certain events, i.e. expected to leave UK, dates of foreign travel
- Educational attainment, current and past
- Gender identity; gender reassignment. NB. Exceptions are surgical procedures
- Geographical locations that are potentially identifiable
- Insurance status, i.e. insurance refused for medical reasons
- Languages, i.e. specific foreign languages spoken, language interpreter needed
- Legal status; court proceedings i.e. lacking mental capacity to make decisions, person granted indefinite leave to remain in the United Kingdom, references to Children Act 1989, Mental Health Act 1983
- Living abodes that are unusual or distinct, e.g. castle
- Personal beliefs, practices and preferences, including: religion, sexuality, choices regarding refusal of clinical intervention or resuscitation, preferences for death and dying
- Personal information, i.e. names, addresses, telephone numbers
- Prisoner of War
- Rare genetic development disorders

- Relationship codes and family events, i.e. marriage, adoption, deaths in family, life crises, extramarital affairs
- Sexual assault
- Sexual orientation

*For specific questions on redacted codes please contact the UK Biobank Access team: access@ukbiobank.ac.uk*