UK Biobank

TPP Primary Care Data for COVID-19 Research

Version 1.1

---

www.ukbiobank.ac.uk
August 2020

This document provides supporting information for the release of primary care data by UK Biobank. **These data are only to be used 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 contains data for approximately 190,000 participants in England covering the primary care (GP) practices with TPP[1] as their data system supplier.

Data from the other main GP data system supplier in England (EMIS[2]), as well as GP data from Scotland and Wales, will be provided in due course. The intention is to update these data approximately every two months.

The primary care data are only to be used for COVID-19 related research. See section 3 for further details.

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 Registering to access primary care data for COVID-19 purposes

If you are a UK Biobank researcher with a project underway, and wish to receive linked health records for COVID-19 research (including primary care data), you can register your interest by logging into the Access Management System (AMS). Go to the 'Data' tab for your project and click the button that says 'Request COVID-19 data and updates'. This takes you to a sign-up website where you can select the specific types of COVID-19 related data you require. The Principal Investigator of the project will need to sign-up to apply for these data and will be asked to confirm that the primary care data will only be used for COVID-19 related research.

If you are a registered researcher but your project has not yet started, please wait until your project has been approved before registering to receive these data. If you are not a registered researcher, please visit the UK Biobank website to learn more about the project and submit a registration. Once registered, you can then submit an application to request these data for COVID-19 related research.

1.3 Downloading primary care data for COVID-19 research purposes

The COVID-19 primary care data can be downloaded via the **Data Portal** on Data Showcase, which will be updated as and when new data becomes available. The tables containing TPP primary care data are:

- covid19_tpp_gp_scripts (drug data)
- covid19_tpp_gp_clinical (clinical coded data)

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

[1] https://www.tpp-uk.com
[2] https://www.emishealth.com

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 is '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 various 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 TPP 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 receives under the COPI regulations are 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 are 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.

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 (jenny.mills@ndph.ox.ac.uk) 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 from TPP 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 hope 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 bear in mind 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.

Although this initial release only contains data from TPP, we have included information on the coding schema used for both TPP and EMIS so that researchers are aware in advance of the key differences. In

particular, it may be useful to download the SNOMED CT look-up tables[3] for EMIS data, as we are not able to provide these directly.

Table 1. System suppliers and their coding classifications

GP System Supplier	Clinical coding classification	Prescription coding classification
TPP	Clinical Terms Version 3 (CTV3 or Read v3) Local TPP codes	Dictionary of Medicines and Devices (dm+d)
EMIS	SNOMED CT Read v2 Local EMIS codes	SNOMED CT

4.2 TPP primary care tables

The TPP clinical and prescription data are provided in two separate tables on the Data Portal, as described below. Please note that 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.

covid19_tpp_gp_clinical

Column name	Description	Encoding
eid	Participant identifier	–
event_dt	Date clinical code was entered	Special codes in Data-Coding 819
code	Clinical code	CTV3: Data-Coding 7128 Local TPP: Data-Coding 8708
code_type	Code type (CTV3[4] or local TPP code)	Data-Coding 3175
value	Value recorded	Special codes in Data-Coding 5702

The clinical code 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

[3] https://isd.digital.nhs.uk/trud3/user/guest/group/0/pack/26

[4] Read codes were updated biannually and distributed under Open Government Licens e via the UKTC Terminology Reference data Update Distribution (TRUD) service - https://isd.digital.nhs.uk/

separately and are not included in this release). Where available, a value field is provided that may give further details. No unit information was available to contextualise the value field.

covid19_tpp_gp_scripts

Column name	Description	Encoding
eid	Participant identifier	–
issue_date	Date prescription was issued	Special codes in Data-Coding 819
dmd_code	dm+d[5]	dm+d codes available from TRUD Special codes in Data-Coding 4214

4.3 Coding systems used in TPP

Each of the coding classification systems used in the TPP data are described below with links to resources for more information. Please note that the volume of coded data in primary care has changed over time and coding practices may be influenced by local procedures, requirements around reporting such as the Quality Outcomes Framework[6] (QOF), and other factors.

Hence, primary care systems are 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.

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) website,[7] however they are scheduled for removal in 2020.[8] As well as the TRUD resources, CTV3 codes are available via the Data Showcase using Data-Coding 7128.

[5] dm+d provides a dictionary of descriptions and codes for medicines and devices used across the NHS.
[6] https://digital.nhs.uk/data-and-information/data-collections-and-data-sets/data-collections/quality-and-outcomes-framework-qof
[7] https://isd.digital.nhs.uk/trud3/user/guest/group/0/pack/9
[8] https://digital.nhs.uk/services/terminology-and-classifications/read-codes

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.

A list of TPP local codes that are present in the current extract and their definitions can be found in Data Showcase Encoding 8708.

4.3.3 dm+d (Dictionary of Medicines and Devices)

The TPP prescription data are coded using dm+d codes (and not the BNF codes that were used in the 2019 release). The dm+d system[9] 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 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. Prescription codes in the TPP data consist of AMP and VMP codes.

[9] https://digital.nhs.uk/data-and-information/information-standards/information-standards-and-data-collections-including-extractions/publications-and-notifications/standards-and-collections/scci0052-dictionary-of-medicines-and-devices-dm-d

Table 4. Example dm+d codes, components and descriptions

dm+d code	dm+d component	Description
109081006	VTM	Metformin
386047000	VMP	Metformin 500mg modified-release tablets
35547511000001101	AMP	Yaltormin SR 500mg tablets (Wockhardt UK Ltd)
8990611000001109	VMPP	Metformin 500mg modified-release tablets 56 tablets
35547911000001108	AMPP	Yaltormin SR 500mg tablets (Wockhardt UK Ltd) 56 tablet

More information about the dm+d model is available as part of a series of short webinars available from the TRUD website (registration required);[10] code lookups are also available. The dm+d codings are provided in XML files; to convert them into tabular format, the dm+d XML Transformation Tool (also downloadable from TRUD) is required. Details on using the dm+d XML Transformation Tool (Windows only) can be found in Appendix A. The NHS Business Services Authority (NHSBSA) also provides a web-based dm+d browser.[11]

4.4 Clinical code look-ups and mapping files

In order to facilitate research on these data, we have compiled clinical code lists for CTV3 and local TPP codes (see Data Showcase encodings 7128 and 8708 respectively). The code_type field in the clinical table indicates which coding system each code belongs to, using Data-Coding 3175. TRUD has historically provided information on how to map from CTV3 to other clinical coding systems, and states that although the final CTV3 release and the Read Browser may no longer be downloadable as Read versions are deprecated and SNOMED CT is adopted, mappings from CTV3 to other coding systems will continue to be available.

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

The accuracy of code lists, definitions and maps should be verified by specialists as part of any analysis undertaken on these data.

[10] https://isd.digital.nhs.uk/trud3/user/authenticated/group/0/pack/6/
[11] https://applications.nhsbsa.nhs.uk/DMDBrowser/DMDBrowser.do

5 Data quality & cleaning

5.1 Data quality

We are making the TPP data available to researchers in a form as close as possible as that provided to UK Biobank by the system supplier. This is to avoid introducing a potential systematic error or bias through data cleaning efforts, for example by removing or altering invalid or erroneous information.

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.

5.2 Data cleaning

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

5.2.1 Data linkage

Data has been provided by TPP to UK Biobank with some limited personal identifiers to enable us to validate the matching algorithm. A small proportion of data (approximately 0.28% of participants included in the TPP data, less than 0.15% of the overall cohort) has not been released due to mismatches with our records.

5.2.2 Redacting potentially disclosive codes

Some clinical codes have the potential to be disclosive – e.g., rare occupation codes. As such, any occupation code that appeared in no more than five records has been redacted. Code descriptions have also been manually reviewed to identify and redact other codes that could be identifying, are related to a person other than the participant such as a family member, or are otherwise deemed to be sensitive and are not clinical in nature.

5.2.3 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. However, there are two circumstances in which no dm+d code is present:

- Some food-related items, bandages, older or obsolete drugs, and some rarer formulations;
- Where an item is mapped to multiple dm+d codes.

In these circumstances, the data has been recoded using Data-Coding 4214.

5.2.4 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 codes are found in Data-Coding 5702.

5.2.5 Dates

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

- 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.

5.2.6 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.

6 Accessing the TPP primary care data

6.1 The Data Portal

The primary care data available for COVID-19 research is only available to download via the Data Portal. This is to enable researchers to access the data as soon as it is released, without the need to add data-fields to a basket and to submit a basket refresh each time new data is available. In order to access the data, you must first register for COVID-19 related data. See Section 1.2 for details on how to do this.

In the Data Showcase, we have generated data-fields 40101 and 40102, corresponding to the COVID-19 clinical and prescription tables, respectively. They display the total number of records in each table. These data-fields are for information only – i.e. researchers cannot add them to a basket.

For guidance how to access and utilise the Data Portal, please see the Accessing Data Guide on the Accessing your data page of Essential Information.

6.2 Sample SQL

Tables on the Data Portal can either be downloaded in full, or queries can be run through the Portal to access specific subsets of data.

Please be aware that due to the size of the tables, some queries may take some time to run and for some more computationally demanding queries, browsers may time-out.

A few simple examples of extracting data-fields using SQL queries on the Portal are given below:

(1) Viewing all the primary care clinical records for participants with CTV3 code 'H060.' (Acute bronchitis (& wheezy)):

```
select * from covid19_tpp_gp_clinical
where code = 'H060.'
```

(2) Counting how many participants have a diagnosis of COVID-19 confirmed by laboratory test recorded in local TPP codes:

```
select count(distinct eid)
from covid19_tpp_gp_clinical
where code = 'Y228d'
```

(3) Viewing the first 100 prescription records where the participant has been prescribed 'Loratadine 5mg/5ml oral solution sugar free' and this has been recorded in dm+d. Note that both VMP and AMP codes are included.

```
select top 100 * from covid19_tpp_gp_scripts
where dmd_code = '36900211000001100'
or dmd_code = '36904111000001102'
```

Note that both CTV3 and dm+d use hierarchy 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 or dm+d Browser (see Section 4.3 for more information). A hierarchy guide for local TPP codes was not available at the time of writing.

Care should be taken with CTV3 codes, since nested within the hierarchy there are sometimes specific codes that indicate a person did not have the condition.

Appendix A. 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).[12]

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.
2. 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.
3. 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.
4. 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 TPP data:
    a. The lookup file for VMPs is "f_vmp_VmpType.csv"
    b. The lookup file for AMPs is "f_amp_AmpType.csv" (this file also serves as a crossmap giving the corresponding VMP for each AMP).

For more information about the components of the dm+d model, including VMPs and AMPs, please see Section 4.3.3.

[12] https://isd.digital.nhs.uk/trud3/user/authenticated/group/0/pack/6/subpack/71/releases