This file is autogenerated. Please do not edit.

TODO - project title

Summary

TODO edit `about.md` to insert a brief lay summary of the work.

Table of contents

Introduction

The aim of this document is to provide full transparency for all parts of the data extraction process. This includes:

- The methodology around how the data extraction process is managed and quality is maintained.
- A full list of all queries used in the extraction, and their associated objectives and assumptions.
- A full list of all clinical codes used for the extraction.

Methodology

After each proposal is approved, a Research Data Engineer (RDE) works closely with the research team to establish precisely what data they require and in what format. The RDE has access to the entire de-identified database and so builds up an expertise as to which projects are feasible and how best to extract the relevant data. The RDE has access to a library of resusable SQL queries for common tasks, and sets of clinical codes for different phenotypes, built up from previous studies. Prior to data extraction, the code is checked and signed off by another RDE.

Reusable queries

This project required the following reusable queries:

- COVID vaccinations
- Lower level super output area

- Index Multiple Deprivation
- Sex
- GET practice and ccg for each patient
- CCG lookup table
- Care home status
- Smoking status
- BMI
- Patient GP encounters
- COVID-related secondary admissions
- Patients with COVID
- Secondary admissions and length of stay
- Secondary discharges
- GET No. LTCS per patient
- Long-term conditions and comorbidities
- Year of birth
- Create table of patients who are registered with a GM GP, and haven't joined the database from June 2022 onwards

Further details for each query can be found below.

COVID vaccinations

To obtain a table with first, second, third... etc vaccine doses per patient.

Assumptions

- GP records can often be duplicated. The assumption is that if a patient receives two vaccines within 14 days of each other then it is likely that both codes refer to the same vaccine.
- The vaccine can appear as a procedure or as a medication. We assume that the presence of either represents a vaccination

Input

```
 Takes two parameters:
        - gp-events-table: string - (table name) the name of the table containing the GP events.
Usually is "RLS.vw_GP_Events" but can be anything with the columns: FK_Patient_Link_ID, EventDate,
and SuppliedCode
        - gp-medications-table: string - (table name) the name of the table containing the GP
medications. Usually is "RLS.vw_GP_Medications" but can be anything with the columns:
FK_Patient_Link_ID, EventDate, and SuppliedCode
```

Output

```
A temp table as follows:
#COVIDVaccinations (FK_Patient_Link_ID, VaccineDate, DaysSinceFirstVaccine)
	- FK_Patient_Link_ID - unique patient id
	- VaccineDose1Date - date of first vaccine (YYYY-MM-DD)
	-	VaccineDose2Date - date of second vaccine (YYYY-MM-DD)
	-	VaccineDose3Date - date of third vaccine (YYYY-MM-DD)
	-	VaccineDose4Date - date of fourth vaccine (YYYY-MM-DD)
	-	VaccineDose5Date - date of fifth vaccine (YYYY-MM-DD)
	-	VaccineDose6Date - date of sixth vaccine (YYYY-MM-DD)
	-	VaccineDose7Date - date of seventh vaccine (YYYY-MM-DD)
```

File: `query-get-covid-vaccines.sql`

Link: https://github.com/rw251/.../query-get-covid-vaccines.sql

---

Lower level super output area

To get the LSOA for each patient.

Assumptions

- Patient data is obtained from multiple sources. Where patients have multiple LSOAs we determine the LSOA as follows:
- If the patients has an LSOA in their primary care data feed we use that as most likely to be up to date
- If every LSOA for a paitent is the same, then we use that
- If there is a single most recently updated LSOA in the database then we use that
- Otherwise the patient's LSOA is considered unknown

Input

```
Assumes there exists a temp table as follows:
#Patients (FK_Patient_Link_ID)
 A distinct list of FK_Patient_Link_IDs for each patient in the cohort
```

Output

```
A temp table as follows:
#PatientLSOA (FK_Patient_Link_ID, LSOA)
	- FK_Patient_Link_ID - unique patient id
	- LSOA_Code - nationally recognised LSOA identifier
```

File: `query-patient-lsoa.sql`

Link: https://github.com/rw251/.../query-patient-lsoa.sql

Index Multiple Deprivation

To get the 2019 Index of Multiple Deprivation (IMD) decile for each patient.

Input

```
Assumes there exists a temp table as follows:
#Patients (FK_Patient_Link_ID)
 A distinct list of FK_Patient_Link_IDs for each patient in the cohort
```

Output

```
A temp table as follows:
#PatientIMDDecile (FK_Patient_Link_ID, IMD2019Decile1IsMostDeprived10IsLeastDeprived)
       - FK_Patient_Link_ID - unique patient id
       - IMD2019Decile1IsMostDeprived10IsLeastDeprived - number 1 to 10 inclusive
```

File: `query-patient-imd.sql`

Link: https://github.com/rw251/.../query-patient-imd.sql

Sex

To get the Sex for each patient.

Assumptions

- Patient data is obtained from multiple sources. Where patients have multiple sexes we determine the sex as follows:
- If the patients has a sex in their primary care data feed we use that as most likely to be up to date
- If every sex for a patient is the same, then we use that
- If there is a single most recently updated sex in the database then we use that
- Otherwise the patient's sex is considered unknown

Input

```
Assumes there exists a temp table as follows:
#Patients (FK_Patient_Link_ID)
 A distinct list of FK_Patient_Link_IDs for each patient in the cohort
```

Output

```
A temp table as follows:
#PatientSex (FK_Patient_Link_ID, Sex)
```

```
        - FK_Patient_Link_ID - unique patient id
        - Sex - M/F
```

File: `query-patient-sex.sql`

Link: https://github.com/rw251/.../query-patient-sex.sql

GET practice and ccg for each patient

For each patient to get the practice id that they are registered to, and the CCG name that the practice belongs to.

Input

```
Assumes there exists a temp table as follows:
#Patients (FK_Patient_Link_ID)
 A distinct list of FK_Patient_Link_IDs for each patient in the cohort
```

Output

```
Two temp tables as follows:
#PatientPractice (FK_Patient_Link_ID, GPPracticeCode)
       - FK_Patient_Link_ID - unique patient id
       - GPPracticeCode - the nationally recognised practice id for the patient
#PatientPracticeAndCCG (FK_Patient_Link_ID, GPPracticeCode, CCG)
       - FK_Patient_Link_ID - unique patient id
       - GPPracticeCode - the nationally recognised practice id for the patient
       - CCG - the name of the patient's CCG
```

File: `query-patient-practice-and-ccg.sql`

Link: https://github.com/rw251/.../query-patient-practice-and-ccg.sql

---

CCG lookup table

To provide lookup table for CCG names. The GMCR provides the CCG id (e.g. '00T', '01G') but not the CCG name. This table can be used in other queries when the output is required to be a ccg name rather than an id.

Input

```
No pre-requisites
```

Output

```
A temp table as follows:
#CCGLookup (CcgId, CcgName)
       - CcgId - Nationally recognised ccg id
       - CcgName - Bolton, Stockport etc..
```

File: `query-ccg-lookup.sql`

Link: https://github.com/rw251/.../query-ccg-lookup.sql

Care home status

To get the care home status for each patient.

Assumptions

- If any of the patient records suggests the patients lives in a care home we will assume that they do

Input

```
Assumes there exists a temp table as follows:
#Patients (FK_Patient_Link_ID)
 A distinct list of FK_Patient_Link_IDs for each patient in the cohort
```

Output

```
A temp table as follows:
#PatientCareHomeStatus (FK_Patient_Link_ID, IsCareHomeResident)
       - FK_Patient_Link_ID - unique patient id
       - IsCareHomeResident - Y/N
```

File: `query-patient-care-home-resident.sql`

Link: https://github.com/rw251/.../query-patient-care-home-resident.sql

---

Smoking status

To get the smoking status for each patient in a cohort.

Assumptions

- We take the most recent smoking status in a patient's record to be correct
- However, there is likely confusion between the "non smoker" and "never smoked" codes. Especially as sometimes the synonyms for these codes overlap. Therefore, a patient wih

a most recent smoking status of "never", but who has previous smoking codes, would be classed as WorstSmokingStatus=non-trivial-smoker / CurrentSmokingStatus=non-smoker

Input

```
 Assumes there exists a temp table as follows:
 #Patients (FK_Patient_Link_ID)
  A distinct list of FK_Patient_Link_IDs for each patient in the cohort
 Also takes one parameter:
        - gp-events-table: string - (table name) the name of the table containing the GP events.
Usually is "RLS.vw_GP_Events" but can be anything with the columns: FK_Patient_Link_ID, EventDate,
FK_Reference_Coding_ID, and FK_Reference_SnomedCT_ID
```

Output

```
 A temp table as follows:
 #PatientSmokingStatus (FK_Patient_Link_ID, PassiveSmoker, WorstSmokingStatus,
CurrentSmokingStatus)
        - FK_Patient_Link_ID - unique patient id
        - PassiveSmoker - Y/N (whether a patient has ever had a code for passive smoking)
        - WorstSmokingStatus - [non-trivial-smoker/trivial-smoker/non-smoker]
        - CurrentSmokingStatus - [non-trivial-smoker/trivial-smoker/non-smoker]
```

File: `query-patient-smoking-status.sql`

Link: https://github.com/rw251/.../query-patient-smoking-status.sql

---

BMI

To get the BMI for each patient in a cohort.

Assumptions

- We take the measurement closest to @IndexDate to be correct

Input

```
 Assumes there exists a temp table as follows:
 #Patients (FK_Patient_Link_ID)
  A distinct list of FK_Patient_Link_IDs for each patient in the cohort
 Also takes one parameter:
        - gp-events-table: string - (table name) the name of the table containing the GP events.
Usually is "RLS.vw_GP_Events" but can be anything with the columns: FK_Patient_Link_ID, EventDate,
FK_Reference_Coding_ID, and FK_Reference_SnomedCT_ID
 Also assumes there is an @IndexDate defined - The index date of the study
```

Output

```
A temp table as follows:
#PatientBMI (FK_Patient_Link_ID, BMI, DateOfBMIMeasurement)
       - FK_Patient_Link_ID - unique patient id
 - BMI
 - DateOfBMIMeasurement
```

File: `query-patient-bmi.sql`

Link: https://github.com/rw251/.../query-patient-bmi.sql

Patient GP encounters

To produce a table of GP encounters for a list of patients. This script uses many codes related to observations (e.g. blood pressure), symptoms, and diagnoses, to infer when GP encounters occured. This script includes face to face and telephone encounters - it will need copying and editing if you don't require both.

Assumptions

- multiple codes on the same day will be classed as one encounter (so max daily encounters per patient is 1)

Input

```
Assumes there exists a temp table as follows:
#Patients (FK_Patient_Link_ID)
 A distinct list of FK_Patient_Link_IDs for each patient in the cohort
```

Output

```
A temp table as follows:
#GPEncounters (FK_Patient_Link_ID, EncounterDate)
       - FK_Patient_Link_ID - unique patient id
       - EncounterDate - date the patient had a GP encounter
```

File: `query-patient-gp-encounters.sql`

Link: https://github.com/rw251/.../query-patient-gp-encounters.sql

---

COVID-related secondary admissions

To classify every admission to secondary care based on whether it is a COVID or non-COVID related. A COVID-related admission is classed as an admission within 4 weeks after, or up to 2 weeks before a positive test.

Input

```
 Takes one parameter
  - start-date: string - (YYYY-MM-DD) the date to count diagnoses from. Usually this should be
2020-01-01.
        -       all-patients: boolean - (true/false) if true, then all patients are included,
otherwise only those in the pre-existing #Patients table.
        - gp-events-table: string - (table name) the name of the table containing the GP events.
Usually is "RLS.vw_GP_Events" but can be anything with the columns: FK_Patient_Link_ID, EventDate,
and SuppliedCode
 And assumes there exists two temp tables as follows:
 #Patients (FK_Patient_Link_ID)
  A distinct list of FK_Patient_Link_IDs for each patient in the cohort
 #Admissions (FK_Patient_Link_ID, AdmissionDate, AcuteProvider)
  A distinct list of the admissions for each patient in the cohort
```

Output

```
 A temp table as follows:
 #COVIDUtilisationAdmissions (FK_Patient_Link_ID, AdmissionDate, AcuteProvider,
CovidHealthcareUtilisation)
        - FK_Patient_Link_ID - unique patient id
        - AdmissionDate - date of admission (YYYY-MM-DD)
        - AcuteProvider - Bolton, SRFT, Stockport etc..
        - CovidHealthcareUtilisation - 'TRUE' if admission within 4 weeks after, or up to 14 days
before, a positive test
```

File: `query-admissions-covid-utilisation.sql`

Link: https://github.com/rw251/.../query-admissions-covid-utilisation.sql

---

Patients with COVID

To get tables of all patients with a COVID diagnosis in their record. This now includes a table that has reinfections. This uses a 90 day cut-off to rule out patients that get multiple tests for a single infection. This 90 day cut-off is also used in the government COVID dashboard. In the first wave, prior to widespread COVID testing, and prior to the correct clinical codes being available to clinicians, infections were recorded in a variety of ways. We therefore take the first diagnosis from any code indicative of COVID. However, for subsequent infections we insist on the presence of a positive COVID test (PCR or antigen) as opposed to simply a diagnosis code. This is to avoid the situation where a hospital diagnosis code gets entered into the primary care record several months after the actual infection.

Input

```
 Takes three parameters
  - start-date: string - (YYYY-MM-DD) the date to count diagnoses from. Usually this should be
2020-01-01.
```

```
        -       all-patients: boolean - (true/false) if true, then all patients are included,
otherwise only those in the pre-existing #Patients table.
        - gp-events-table: string - (table name) the name of the table containing the GP events.
Usually is "SharedCare.GP_Events" but can be anything with the columns: FK_Patient_Link_ID,
EventDate, and SuppliedCode
```

Output

```
Three temp tables as follows:
#CovidPatients (FK_Patient_Link_ID, FirstCovidPositiveDate)
        - FK_Patient_Link_ID - unique patient id
        - FirstCovidPositiveDate - earliest COVID diagnosis
#CovidPatientsAllDiagnoses (FK_Patient_Link_ID, CovidPositiveDate)
        - FK_Patient_Link_ID - unique patient id
        - CovidPositiveDate - any COVID diagnosis
#CovidPatientsMultipleDiagnoses
        -       FK_Patient_Link_ID - unique patient id
        -       FirstCovidPositiveDate - date of first COVID diagnosis
        -       SecondCovidPositiveDate - date of second COVID diagnosis
        -       ThirdCovidPositiveDate - date of third COVID diagnosis
        -       FourthCovidPositiveDate - date of fourth COVID diagnosis
        -       FifthCovidPositiveDate - date of fifth COVID diagnosis
```

File: `query-patients-with-covid.sql`

Link: https://github.com/rw251/.../query-patients-with-covid.sql

---

Secondary admissions and length of stay

To obtain a table with every secondary care admission, along with the acute provider, the date of admission, the date of discharge, and the length of stay.

Input

```
One parameter
        -       all-patients: boolean - (true/false) if true, then all patients are included,
otherwise only those in the pre-existing #Patients table.
```

Output

```
Two temp table as follows:
#Admissions (FK_Patient_Link_ID, AdmissionDate, AcuteProvider)
        - FK_Patient_Link_ID - unique patient id
        - AdmissionDate - date of admission (YYYY-MM-DD)
        - AcuteProvider - Bolton, SRFT, Stockport etc..
  (Limited to one admission per person per hospital per day, because if a patient has 2 admissions
   on the same day to the same hopsital then it's most likely data duplication rather than two
short
   hospital stays)
#LengthOfStay (FK_Patient_Link_ID, AdmissionDate)
        - FK_Patient_Link_ID - unique patient id
```

```
        - AdmissionDate - date of admission (YYYY-MM-DD)
        - AcuteProvider - Bolton, SRFT, Stockport etc..
        - DischargeDate - date of discharge (YYYY-MM-DD)
        - LengthOfStay - Number of days between admission and discharge. 1 = [0,1) days, 2 = [1,2)
days, etc.
```

File: `query-get-admissions-and-length-of-stay.sql`

Link: https://github.com/rw251/.../query-get-admissions-and-length-of-stay.sql

---

Secondary discharges

To obtain a table with every secondary care discharge, along with the acute provider, and the date of discharge.

Input

```
 One parameter
        -       all-patients: boolean - (true/false) if true, then all patients are included,
otherwise only those in the pre-existing #Patients table.
```

Output

```
 A temp table as follows:
 #Discharges (FK_Patient_Link_ID, DischargeDate, AcuteProvider)
        - FK_Patient_Link_ID - unique patient id
        - DischargeDate - date of discharge (YYYY-MM-DD)
        - AcuteProvider - Bolton, SRFT, Stockport etc..
  (Limited to one discharge per person per hospital per day, because if a patient has 2 discharges
   on the same day to the same hopsital then it's most likely data duplication rather than two
short
   hospital stays)
```

File: `query-get-discharges.sql`

Link: https://github.com/rw251/.../query-get-discharges.sql

GET No. LTCS per patient

To get the number of long-term conditions for each patient.

Input

```
 Assumes there exists a temp table as follows:
 #PatientsWithLTCs (FK_Patient_Link_ID, LTC)
 Therefore this is run after query-patient-ltcs.sql
```

Output

```
A temp table with a row for each patient with the number of LTCs they have
#NumLTCs (FK_Patient_Link_ID, NumberOfLTCs)
```

File: `query-patient-ltcs-number-of.sql`

Link: https://github.com/rw251/.../query-patient-ltcs-number-of.sql

---

Long-term conditions and comorbidities

To get long-term conditions for each patient within a date range.

Input

```
Assumes there exists a temp table as follows:
#Patients (FK_Patient_Link_ID)
A distinct list of FK_Patient_Link_IDs for each patient in the cohort
@IndexDate - The index date of the study
@MinDate - The minimum date that the study has access from
```

Output

```
A temp table with a row for each patient and ltc combo
#PatientsWithLTCs
  FK_Patient_Link_ID,
  Condition,
  FirstDate,
  LastDate,
  NoInLastYear,number of times mentioned in year prior to index date
  LTCflag, flag to indicate whether is a long condition (Y) or not (N) - see comments on code for
defining long term conditions.
```

File: `query-patient-ltcs-date-range.sql`

Link: https://github.com/rw251/.../query-patient-ltcs-date-range.sql

---

Year of birth

To get the year of birth for each patient.

Assumptions

- Patient data is obtained from multiple sources. Where patients have multiple YOBs we determine the YOB as follows:

- If the patients has a YOB in their primary care data feed we use that as most likely to be up to date
- If every YOB for a patient is the same, then we use that
- If there is a single most recently updated YOB in the database then we use that
- Otherwise we take the highest YOB for the patient that is not in the future

Input

```
Assumes there exists a temp table as follows:
#Patients (FK_Patient_Link_ID)
 A distinct list of FK_Patient_Link_IDs for each patient in the cohort
```

Output

```
A temp table as follows:
#PatientYearOfBirth (FK_Patient_Link_ID, YearOfBirth)
       - FK_Patient_Link_ID - unique patient id
       - YearOfBirth - INT
```

File: `query-patient-year-of-birth.sql`

Link: https://github.com/rw251/.../query-patient-year-of-birth.sql

---

Create table of patients who are registered with a GM GP, and haven't joined the database from June 2022 onwards

undefined

Input

```
undefined
```

Output

```
undefined
```

File: `query-get-possible-patients.sql`

Link: https://github.com/rw251/.../query-get-possible-patients.sql

Clinical code sets

This project required the following clinical code sets:

- diabetes-type-i v1
- diabetes-type-ii v1
- covid-positive-antigen-test v1
- covid-positive-pcr-test v1
- covid-positive-test-other v1
- bnf-gastro-intestinal-meds v1
- bnf-cardiovascular-meds v1
- bnf-respiratory-meds v1
- bnf-cns-meds v1
- bnf-infections-meds v1
- bnf-endocrine-meds v1
- bnf-obstetrics-gynaecology-meds v1
- bnf-malignant-disease-immunosuppression-meds v1
- bnf-nutrition-bloods-meds v1
- bnf-muskuloskeletal-joint-meds v1
- bnf-eye-meds v1
- bnf-ear-nose-throat-meds v1
- bnf-skin-meds v1
- bnf-immunological-meds v1
- bnf-anaesthesia-meds v1
- sglt2-inhibitors v1
- metformin v1
- insulin v1
- ace-inhibitor v2
- angiotensin-receptor-blockers v1
- aspirin v1
- clopidogrel v1
- alogliptin v1
- linagliptin v1
- saxagliptin v1
- sitagliptin v1
- vildagliptin v1
- systolic-blood-pressure v1
- diastolic-blood-pressure v1
- hba1c v2
- cholesterol v2
- ldl-cholesterol v1
- hdl-cholesterol v1
- triglycerides v1
- egfr v1
- height v1
- weight v1

- bmi v2
- smoking-status-current v1
- smoking-status-currently-not v1
- smoking-status-ex v1
- smoking-status-ex-trivial v1
- smoking-status-never v1
- smoking-status-passive v1
- smoking-status-trivial v1
- covid-vaccination v1

Further details for each code set can be found below.

Diabetes mellitus type 1

Any diagnosis of T1DM. A super set of the QOF business rule.

Developed from https://getset.ga.

Prevalence log

By examining the prevalence of codes (number of patients with the code in their record) broken down by clinical system, we can attempt to validate the clinical code sets and the reporting of the conditions. Here is a log for this code set. The prevalence range `0.42% - 0.48%` suggests that this code set is well defined.

Date	Practice system	Population	Patients from ID	Patient from code
2021-05-07	EMIS	2605681	11381 (0.44%)	11381 (0.44%)
2021-05-07	TPP	210817	887 (0.42%)	887 (0.42%)
2021-05-07	Vision	334632	1607 (0.48%)	1607 (0.48%)

LINK: https://github.com/rw251/.../conditions/diabetes-type-i/1

Diabetes mellitus type 2

Any diagnosis of T2DM. A super set of the QOF business rule. Includes "adult onset" diabetes, but DOES NOT include "maturity onset" diabetes.

Developed from https://getset.ga.

Prevalence log

By examining the prevalence of codes (number of patients with the code in their record) broken down by clinical system, we can attempt to validate the clinical code sets and the

reporting of the conditions. Here is a log for this code set. The prevalence range `5.06% - 5.20%` suggests that this code set is well defined.

Date	Practice system	Population	Patients from ID	Patient from code
2021-05-07	EMIS	2605681	133938 (5.14%)	133938 (5.14%)
2021-05-07	TPP	210817	10954 (5.20%)	10954 (5.20%)
2021-05-07	Vision	334632	16936 (5.06%)	16933 (5.06%)

LINK: https://github.com/rw251/.../conditions/diabetes-type-ii/1

COVID-19 positive antigen test

A code that indicates that a person has a positive antigen test for COVID-19.

COVID positive tests in primary care

The codes used in primary care to indicate a positive COVID test can be split into 3 types: antigen test, PCR test and other. We keep these as separate code sets. However due to the way that COVID diagnoses are recorded in different ways in different GP systems, and because some codes are ambiguous, currently it only makes sense to group these 3 code sets together. Therefore the prevalence log below is for the combined code sets of `covid-positive-antigen-test`, `covid-positive-pcr-test` and `covid-positive-test-other`.

Prevalence log

By examining the prevalence of codes (number of patients with the code in their record) broken down by clinical system, we can attempt to validate the clinical code sets and the reporting of the conditions. Here is a log for this code set. The prevalence range `18.6% - 20.5%` suggests that this code set is likely well defined. *NB - this code set needs to rely on the SuppliedCode in the database rather than the foreign key ids.*

Date	Practice system	Population	Patients from ID	Patient from code
2022-02-25	EMIS	2656041	152972 (5.76%)	545759 (20.5%)
2022-02-25	TPP	212453	256 (0.12%)	39503 (18.6%)
2022-02-25	Vision	341354	9440 (2.77%)	65963 (19.3%)

LINK: https://github.com/rw251/.../tests/covid-positive-antigen-test/1

COVID-19 positive pcr test

A code that indicates that a person has a positive pcr test for COVID-19.

COVID positive tests in primary care

The codes used in primary care to indicate a positive COVID test can be split into 3 types: antigen test, PCR test and other. We keep these as separate code sets. However due to the way that COVID diagnoses are recorded in different ways in different GP systems, and because some codes are ambiguous, currently it only makes sense to group these 3 code sets together. Therefore the prevalence log below is for the combined code sets of `covid-positive-antigen-test`, `covid-positive-pcr-test` and `covid-positive-test-other`.

Prevalence log

By examining the prevalence of codes (number of patients with the code in their record) broken down by clinical system, we can attempt to validate the clinical code sets and the reporting of the conditions. Here is a log for this code set. The prevalence range `18.6% - 20.5%` suggests that this code set is likely well defined. *NB - this code set needs to rely on the SuppliedCode in the database rather than the foreign key ids.*

Date	Practice system	Population	Patients from ID	Patient from code
2022-02-25	EMIS	2656041	152972 (5.76%)	545759 (20.5%)
2022-02-25	TPP	212453	256 (0.12%)	39503 (18.6%)
2022-02-25	Vision	341354	9440 (2.77%)	65963 (19.3%)

LINK: https://github.com/rw251/.../tests/covid-positive-pcr-test/1

COVID-19 positive test - other

A code that indicates that a person has a positive test for COVID-19, but where the type of test (antigen or PCR) is unknown.

COVID positive tests in primary care

The codes used in primary care to indicate a positive COVID test can be split into 3 types: antigen test, PCR test and other. We keep these as separate code sets. However due to the way that COVID diagnoses are recorded in different ways in different GP systems, and because some codes are ambiguous, currently it only makes sense to group these 3 code sets together. Therefore the prevalence log below is for the combined code sets of `covid-positive-antigen-test`, `covid-positive-pcr-test` and `covid-positive-test-other`.

Prevalence log

By examining the prevalence of codes (number of patients with the code in their record) broken down by clinical system, we can attempt to validate the clinical code sets and the reporting of the conditions. Here is a log for this code set. The prevalence range

`18.6% - 20.5%` suggests that this code set is likely well defined. *NB - this code set needs to rely on the SuppliedCode in the database rather than the foreign key ids.*

Date	Practice system	Population	Patients from ID	Patient from code
2022-02-25	EMIS	2656041	152972 (5.76%)	545759 (20.5%)
2022-02-25	TPP	212453	256 (0.12%)	39503 (18.6%)
2022-02-25	Vision	341354	9440 (2.77%)	65963 (19.3%)

LINK: https://github.com/rw251/.../tests/covid-positive-test-other/1

Gastrointestinal medications

This code set was based on a list of SNOMED codes, derived from BNF Chapter 1 codes.

Prevalence log

By examining the prevalence of codes (number of patients with the code in their record) broken down by clinical system, we can attempt to validate the clinical code sets and the reporting of the conditions. Here is a log for this code set. The prevalence range `51.26% - 59.02%` suggests that this code set is well defined.

Date	Practice system	Population	Patients from ID	Patient from code
2022-06-28	EMIS	2662570	1366098 (51.26%)	1490771 (55.94%)
2022-06-28	TPP	212696	125665 (59.02%)	76008 (35.71%)
2022-06-28	Vision	342344	180744 (52.67%)	159497 (46.48%)

LINK: https://github.com/rw251/.../medications/bnf-gastro-intestinal-meds/1

Cardiovascular medications

This code set was based on a list of SNOMED codes, derived from BNF codes, provided by the study team for RQ32.

Prevalence log

By examining the prevalence of codes (number of patients with the code in their record) broken down by clinical system, we can attempt to validate the clinical code sets and the reporting of the conditions. Here is a log for this code set. The prevalence range `30.2% - 31.4%` suggests that this code set is well defined.

Date	Practice system	Population	Patients from ID	Patient from code
2021-11-08	EMIS	2641576	811072 (30.70%)	810948 (30.70%)
2021-11-08	TPP	211880	66547 (31.41%)	66557 (31.41%)
2021-11-08	Vision	339069	102504 (30.23%)	102504 (30.23%)

LINK: https://github.com/rw251/.../medications/bnf-cardiovascular-meds/1

Respiratory medications

This code set was based on a list of SNOMED codes, derived from BNF Chapter 3 codes.

Prevalence log

By examining the prevalence of codes (number of patients with the code in their record) broken down by clinical system, we can attempt to validate the clinical code sets and the reporting of the conditions. Here is a log for this code set. The prevalence range `48.00% - 51.56%` suggests that this code set is well defined.

Date	Practice system	Population	Patients from ID	Patient from code
2022-06-08	EMIS	2664831	1183167 (44.40%)	1314730 (49.34%)
2022-06-08	TPP	212907	113507 (53.31%)	109769 (51.56%)
2022-06-08	Vision	343146	156589 (45.63%)	164688 (48.00%)

LINK: https://github.com/rw251/.../medications/bnf-respiratory-meds/1

Central Nervous System medications

This code set was based on a list of SNOMED codes, derived from BNF codes, provided by the study team for RQ32.

Prevalence log

NEED TO UPDATE BELOW

By examining the prevalence of codes (number of patients with the code in their record) broken down by clinical system, we can attempt to validate the clinical code sets and the reporting of the conditions. Here is a log for this code set. The prevalence range `56% - 64%` suggests that this code set is reasonably well defined.

Date	Practice system	Population	Patients from ID	Patient from code
2021-11-08	EMIS	2641576	1714809 (64.92%)	1714733 (64.91%)
2021-11-08	TPP	211880	129171 (60.96%)	129196 (60.98%)
2021-11-08	Vision	339069	18966 (56.03%)	18965 (56.03%)

LINK: https://github.com/rw251/.../medications/bnf-cns-meds/1

Infections medications

This code set was based on a list of SNOMED codes, derived from BNF Chapter 5 codes.

Prevalence log

By examining the prevalence of codes (number of patients with the code in their record) broken down by clinical system, we can attempt to validate the clinical code sets and the reporting of the conditions. Here is a log for this code set. The prevalence range `70.83% - 73.72%` suggests that this code set is well defined.

Date	Practice system	Population	Patients from ID	Patient from code
2022-06-08	EMIS	2664831	1759250 (66.02%)	1964444 (73.72%)
2022-06-08	TPP	212907	169942 (79.82%)	151597 (71.20%)
2022-06-08	Vision	343146	237233 (69.13%)	243057 (70.83%)

LINK: https://github.com/rw251/.../medications/bnf-infections-meds/1

Endocrine medications

This code set was based on a list of SNOMED codes, derived from BNF codes, provided by the study team for RQ32.

Prevalence log

By examining the prevalence of codes (number of patients with the code in their record) broken down by clinical system, we can attempt to validate the clinical code sets and the reporting of the conditions. Here is a log for this code set. The prevalence range `26.3% - 27.3%` suggests that this code set is well defined.

Date	Practice system	Population	Patients from ID	Patient from code
2021-11-09	EMIS	2642853	696800 (26.37%)	695807 (26.33%)
2021-11-09	TPP	211895	57804 (27.28%)	57801 (27.28%)
2021-11-09	Vision	339130	91061 (26.85%)	91037 (26.84%)
LINK: https://github.com/rw251/.../medications/bnf-endocrine-meds/1				

Obstetrics, gynaecology and urinary tract disorders medications

This code set was based on a list of SNOMED codes, derived from BNF Chapter 5 codes.

Prevalence log

By examining the prevalence of codes (number of patients with the code in their record) broken down by clinical system, we can attempt to validate the clinical code sets and the reporting of the conditions. Here is a log for this code set. The prevalence range `68.05% - 71.00%` suggests that this code set is well defined.

Date	Practice system	Population	Patients from ID	Patient from code
2022-06-29	EMIS	2664831	1642795 (61.65%)	1832397 (68.76%)
2022-06-29	TPP	212907	157636 (74.04%)	151165 (71.00%)
2022-06-29	Vision	343146	222254 (64.77%)	233527 (68.05%)
LINK: https://github.com/rw251/.../medications/bnf-obstetrics-gynaecology-meds/1				

Malignant disease and immunosuppression medications

This code set was based on a list of SNOMED codes, derived from BNF Chapter 8 codes.

Prevalence log

By examining the prevalence of codes (number of patients with the code in their record) broken down by clinical system, we can attempt to validate the clinical code sets and the reporting of the conditions. Here is a log for this code set. The prevalence range `1.66% - 1.78%` suggests that this code set is well defined.

Date	Practice system	Population	Patients from ID	Patient from code
2022-06-08	EMIS	2664831	46054 (1.73%)	48167 (1.81%)
2022-06-08	TPP	212907	3797 (1.78%)	2857 (1.34%)
2022-06-08	Vision	343146	5692 (1.66%)	5074 (1.47%)

LINK: https://github.com/rw251/.../medications/bnf-malignant-disease-immunosuppression-meds/1

Nutrition and blood medications

This code set was based on a list of SNOMED codes, derived from BNF Chapter 9 codes.

Prevalence log

By examining the prevalence of codes (number of patients with the code in their record) broken down by clinical system, we can attempt to validate the clinical code sets and the reporting of the conditions. Here is a log for this code set. The prevalence range `33.31% - 37.48%` suggests that this code set is well defined.

Date	Practice system	Population	Patients from ID	Patient from code
2022-06-29	EMIS	2664831	882352 (33.11%)	953268 (35.77%)
2022-06-29	TPP	212907	74521 (35.00%)	61290 (28.79%)
2022-06-29	Vision	343146	108654 (31.66%)	105945 (30.87%)

LINK: https://github.com/rw251/.../medications/bnf-nutrition-bloods-meds/1

Muskuloskeletal and joint medications

This code set was based on a list of SNOMED codes, derived from BNF Chapter 10 codes.

Prevalence log

By examining the prevalence of codes (number of patients with the code in their record) broken down by clinical system, we can attempt to validate the clinical code sets and the reporting of the conditions. Here is a log for this code set. The prevalence range `46.91% - 53.11%` suggests that this code set is well defined.

Date	Practice system	Population	Patients from ID	Patient from code
2022-06-29	EMIS	2664831	1214181 (45.56%)	1325568 (49.74%)
2022-06-29	TPP	212907	118324 (55.58%)	113077 (53.11%)
2022-06-29	Vision	343146	153929 (44.86%)	160958 (46.91%)

LINK: https://github.com/rw251/.../medications/bnf-muskuloskeletal-joint-meds/1

Eye medications

This code set was based on a list of SNOMED codes, derived from BNF codes for chapter 11.

Prevalence log

By examining the prevalence of codes (number of patients with the code in their record) broken down by clinical system, we can attempt to validate the clinical code sets and the reporting of the conditions. Here is a log for this code set. The prevalence range `29.72% - 34.93%` suggests that this code set is well defined.

Date	Practice system	Population	Patients from ID	Patient from code
2022-06-28	EMIS	2664831	826022 (31.00%)	912789 (34.25%)
2022-06-28	TPP	212907	74373 (34.93%)	69551 (32.67%)
2022-06-28	Vision	343146	101986 (29.72%)	106901 (31.15%)

LINK: https://github.com/rw251/.../medications/bnf-eye-meds/1

Ear, nose and throat medications

This code set was based on a list of SNOMED codes, derived from BNF codes for chapter 12.

Prevalence log

By examining the prevalence of codes (number of patients with the code in their record) broken down by clinical system, we can attempt to validate the clinical code sets and the reporting of the conditions. Here is a log for this code set. The prevalence range `31.67% - 36.05%` suggests that this code set is well defined.

Date	Practice system	Population	Patients from ID	Patient from code
2022-06-28	EMIS	2664831	961083 (36.05%)	1042189 (39.11%)
2022-06-28	TPP	212907	67420 (31.67%)	34777 (16.33%)
2022-06-28	Vision	343146	112365 (32.75%)	104120 (30.34%)

LINK: https://github.com/rw251/.../medications/bnf-ear-nose-throat-meds/1

Skin medications

This code set was based on a list of SNOMED codes, derived from BNF codes for chapter 13.

Prevalence log

update:

By examining the prevalence of codes (number of patients with the code in their record) broken down by clinical system, we can attempt to validate the clinical code sets and the reporting of the conditions. Here is a log for this code set. The prevalence range `30.2% - 31.4%` suggests that this code set is well defined.

Date	Practice system	Population	Patients from ID	Patient from code
2021-11-08	EMIS	2641576	811072 (30.70%)	810948 (30.70%)
2021-11-08	TPP	211880	66547 (31.41%)	66557 (31.41%)
2021-11-08	Vision	339069	102504 (30.23%)	102504 (30.23%)

LINK: https://github.com/rw251/.../medications/bnf-skin-meds/1

Skin medications

This code set was based on a list of SNOMED codes, derived from BNF codes for chapter 13.

Prevalence log

By examining the prevalence of codes (number of patients with the code in their record) broken down by clinical system, we can attempt to validate the clinical code sets and the reporting of the conditions. Here is a log for this code set. The prevalence range `15.15% - 59.72%` suggests that there are missing codes from one or more code set.

Date	Practice system	Population	Patients from ID	Patient from code
2022-06-28	EMIS	2664831	1591464 (59.72%)	1705445 (64.00%)
2022-06-28	TPP	212907	32249 (15.15%)	31720 (14.90%)
2022-06-28	Vision	343146	96281 (28.06%)	104339 (30.41%)

LINK: https://github.com/rw251/.../medications/bnf-immunological-meds/1

Anaesthesia medications

This code set was based on a list of SNOMED codes, derived from BNF codes for chapter 13.

Prevalence log

By examining the prevalence of codes (number of patients with the code in their record) broken down by clinical system, we can attempt to validate the clinical code sets and the reporting of the conditions. Here is a log for this code set. The prevalence range `1.31% - 8.07%` suggests potential missing codes.

Date	Practice system	Population	Patients from ID	Patient from code
2022-06-28	EMIS	2664831	215101 (8.07%)	229808 (8.62%)
2022-06-28	TPP	211880	2798 (1.31%)	1411 (0.66%)
2022-06-28	Vision	339069	22006 (6.41%)	23622 (6.88%)

LINK: https://github.com/rw251/.../medications/bnf-anaesthesia-meds/1

SGLT2 inhibitors (gliflozins)

Any prescription of a SGLT2 inhibitor (gliflozin).

Prevalence log

By examining the prevalence of codes (number of patients with the code in their record) broken down by clinical system, we can attempt to validate the clinical code sets and the reporting of the conditions. Here is a log for this code set. The prevalence range `1.25% - 1.25%` for EMIS and Vision suggests that this code set is well defined. The figure of `0.90%` for TPP is lower than expected, but TPP have the smallest patient population so a degree of variability is to be expected.

Date	Practice system	Population	Patients from ID	Patient from code
2021-08-26	EMIS	2623304	32817 (1.25%)	32817 (1.25%)
2021-08-26	TPP	211610	1899 (0.90%)	1899 (0.90%)
2021-08-26	Vision	337028	4211 (1.25%)	4211 (1.25%)

LINK: https://github.com/rw251/.../medications/sglt2-inhibitors/1

Metformin

This code set was originally created for the SMASH safe medication dashboard and has been validated in practice.

Prevalence log

By examining the prevalence of codes (number of patients with the code in their record) broken down by clinical system, we can attempt to validate the clinical code sets and the reporting of the conditions. Here is a log for this code set. The prevalence range `4.75% - 5.19%` suggests that this code set is well defined.

Date	Practice system	Population	Patients from ID	Patient from code
2021-05-07	EMIS	2605681	135082 (5.18%)	135136 (5.19%)
2021-05-07	TPP	210817	10016 (4.75%)	10016 (4.75%)
2021-05-07	Vision	334632	16809 (5.02%)	16809 (5.02%)

LINK: https://github.com/rw251/.../medications/metformin/1

Insulin

Any code representing a prescription (or drug) of insulin.

Prevalence log

By examining the prevalence of codes (number of patients with the code in their record) broken down by clinical system, we can attempt to validate the clinical code sets and the reporting of the conditions. Here is a log for this code set. The prevalence range `1.21% - 1.32%` suggests that this code set is well defined.

Date	Practice system	Population	Patients from ID	Patient from code
2021-08-26	EMIS	2623304	34737 (1.32%)	34737 (1.32%)
2021-08-26	TPP	211610	2573 (1.22%)	2573 (1.22%)
2021-08-26	Vision	337028	4071 (1.21%)	4071 (1.21%)

LINK: https://github.com/rw251/.../medications/insulin/1

ACE Inhibitors

Any code for a prescription of an ACE Inhibitor.

NB This does not include ARBs. Please use v1 of this code set for a combined ACEI/ARB code set

Prevalence log

By examining the prevalence of codes (number of patients with the code in their record) broken down by clinical system, we can attempt to validate the clinical code sets and the

reporting of the conditions. Here is a log for this code set. The prevalence range `11.00% - 11.82%` suggests that this code set is well defined.

Date	Practice system	Population	Patients from ID	Patient from code
2021-08-26	EMIS	2623304	289565 (11.04%)	289582 (11.04%)
2021-08-26	TPP	211610	25005 (11.82%)	25005 (11.82%)
2021-08-26	Vision	337028	37078 (11.00%)	37078 (11.00%)

LINK: https://github.com/rw251/.../medications/ace-inhibitor/2

Angiotensin receptor blockers (ARBs)

Any code for a prescription of an angiotensin receptor blocker (ARB).

NB This does not include ACEIs. Please use v1 of the ACEI code set for a combined ACEI/ARB code set

Prevalence log

By examining the prevalence of codes (number of patients with the code in their record) broken down by clinical system, we can attempt to validate the clinical code sets and the reporting of the conditions. Here is a log for this code set. The prevalence range `4.43% - 4.80%` suggests that this code set is well defined.

Date	Practice system	Population	Patients from ID	Patient from code
2021-08-26	EMIS	2623304	117732 (4.49%)	117727 (4.49%)
2021-08-26	TPP	211610	10150 (4.80%)	10150 (4.80%)
2021-08-26	Vision	337028	14934 (4.43%)	14934 (4.43%)

LINK: https://github.com/rw251/.../medications/angiotensin-receptor-blockers/1

Aspirin

This code set was originally created for the SMASH safe medication dashboard and has been validated in practice.

Prevalence log

By examining the prevalence of codes (number of patients with the code in their record) broken down by clinical system, we can attempt to validate the clinical code sets and the reporting of the conditions. Here is a log for this code set. The prevalence range `8.45% - 8.76%` suggests that this code set is well defined.

Date	Practice system	Population	Patients from ID	Patient from code
2021-05-07	EMIS	2605681	228074 (8.75%)	228315 (8.76%)
2021-05-07	TPP	210817	18418 (8.74%)	18418 (8.74%)
2021-05-07	Vision	334632	28276 (8.45%)	28276 (8.45%)

LINK: https://github.com/rw251/.../medications/aspirin/1

Clopidogrel

This code set was originally created for the SMASH safe medication dashboard and has been validated in practice.

Prevalence log

By examining the prevalence of codes (number of patients with the code in their record) broken down by clinical system, we can attempt to validate the clinical code sets and the reporting of the conditions. Here is a log for this code set. The prevalence range `2.62% - 2.86%` suggests that this code set is well defined.

Date	Practice system	Population	Patients from ID	Patient from code
2021-05-07	EMIS	2605681	72859 (2.80%)	72859 (2.80%)
2021-05-07	TPP	210817	5515 (2.62%)	5515 (2.62%)
2021-05-07	Vision	334632	9568 (2.86%)	9568 (2.86%)

LINK: https://github.com/rw251/.../medications/clopidogrel/1

Alogliptin

Category: "antidiabetics".

Any prescription of alogliptin.

This code set has been autogenerated by identifying the relevant codes in the BNF files. The BNF codes are then mapped to SNOMED, and then on to EMIS, Readv2 and CTV3.

Prevalence log

By examining the prevalence of codes (number of patients with the code in their record) broken down by clinical system, we can attempt to validate the clinical code sets and the reporting of the conditions. Here is a log for this code set. The prevalence range `0.58% - 0.91%` suggests that this code set is perhaps well defined, though the value for Vision practices is a bit high.

Date	Practice system	Population	Patients from ID	Patient from code
2022-06-28	EMIS	2664831	17696 (0.66%)	18433 (0.69%)
2022-06-28	TPP	212907	1235 (0.58%)	1518 (0.71%)
2022-06-28	Vision	343146	3113 (0.91%)	3135 (0.91%)

LINK: https://github.com/rw251/.../medications/alogliptin/1

Linagliptin

Category: "antidiabetics".

Any prescription of linagliptin.

This code set has been autogenerated by identifying the relevant codes in the BNF files. The BNF codes are then mapped to SNOMED, and then on to EMIS, Readv2 and CTV3.

Prevalence log

By examining the prevalence of codes (number of patients with the code in their record) broken down by clinical system, we can attempt to validate the clinical code sets and the reporting of the conditions. Here is a log for this code set. The prevalence range `0.23% - 0.55%` suggests that this code set is not well defined.

Date	Practice system	Population	Patients from ID	Patient from code
2022-06-28	EMIS	2664831	11065 (0.42%)	11731 (0.44%)
2022-06-28	TPP	212907	1167 (0.55%)	1392 (0.65%)
2022-06-28	Vision	343146	774 (0.23%)	778 (0.23%)

LINK: https://github.com/rw251/.../medications/linagliptin/1

Saxagliptin

Category: "antidiabetics".

Any prescription of saxagliptin.

This code set has been autogenerated by identifying the relevant codes in the BNF files. The BNF codes are then mapped to SNOMED, and then on to EMIS, Readv2 and CTV3.

Prevalence log

By examining the prevalence of codes (number of patients with the code in their record) broken down by clinical system, we can attempt to validate the clinical code sets and the

reporting of the conditions. Here is a log for this code set. The prevalence range `0.06% - 0.19%` suggests that this code set is not well defined.

Date	Practice system	Population	Patients from ID	Patient from code
2022-06-28	EMIS	2664831	3549 (0.13%)	3706 (0.14%)
2022-06-28	TPP	212907	395 (0.19%)	413 (0.19%)
2022-06-28	Vision	343146	223 (0.06%)	226 (0.07%)

LINK: https://github.com/rw251/.../medications/saxagliptin/1

Sitagliptin

Category: "antidiabetics".

Any prescription of sitagliptin.

This code set has been autogenerated by identifying the relevant codes in the BNF files. The BNF codes are then mapped to SNOMED, and then on to EMIS, Readv2 and CTV3.

Prevalence log

By examining the prevalence of codes (number of patients with the code in their record) broken down by clinical system, we can attempt to validate the clinical code sets and the reporting of the conditions. Here is a log for this code set. The prevalence range `0.59% - 0.91%` suggests that this code set is not well defined.

Date	Practice system	Population	Patients from ID	Patient from code
2022-06-28	EMIS	2664831	20007 (0.75%)	21138 (0.79%)
2022-06-28	TPP	212907	1937 (0.91%)	2224 (1.04%)
2022-06-28	Vision	343146	2015 (0.59%)	2028 (0.59%)

LINK: https://github.com/rw251/.../medications/sitagliptin/1

Vildagliptin

Category: "antidiabetics".

Any prescription of vildagliptin.

This code set has been autogenerated by identifying the relevant codes in the BNF files. The BNF codes are then mapped to SNOMED, and then on to EMIS, Readv2 and CTV3.

Prevalence log

By examining the prevalence of codes (number of patients with the code in their record) broken down by clinical system, we can attempt to validate the clinical code sets and the reporting of the conditions. Here is a log for this code set. The prevalence range `0.06% - 0.10%` suggests that this code set is well defined.

Date	Practice system	Population	Patients from ID	Patient from code
2022-06-28	EMIS	2664831	1829 (0.07%)	1919 (0.07%)
2022-06-28	TPP	212907	207 (0.10%)	217 (0.10%)
2022-06-28	Vision	343146	200 (0.06%)	200 (0.06%)

LINK: https://github.com/rw251/.../medications/vildagliptin/1

Systolic Blood pressure

Any indication that systolic blood pressure has been recorded for a patient. This code set only includes codes that are accompanied by a value (`2469. - O/E - Systolic BP reading`).

Blood pressure codes retrieved from GetSet and metadata available in this directory.

Prevalence log

By examining the prevalence of codes (number of patients with the code in their record) broken down by clinical system, we can attempt to validate the clinical code sets and the reporting of the conditions. Here is a log for this code set. The prevalence range `64.46% - 67.00%` suggests that this code set is likely well defined.

Date	Practice system	Population	Patients from ID	Patient from code
2021-10-13	EMIS	26929848	1741342 (66.21%)	1741342 (66.21%)
2021-10-13	TPP	211812	137571 (64.95%)	137571 (64.95%)
2021-10-13	Vision	338205	208971 (61.79%)	208971 (61.79%)
LINK: https://github.com/rw251/.../tests/systolic-blood-pressure/1				

Diastolic Blood pressure

Any diastolic blood pressure measurements, with values, that have been recorded for a patient.

Blood pressure codes retrieved from GetSet and metadata available in this directory.

Prevalence log

By examining the prevalence of codes (number of patients with the code in their record) broken down by clinical system, we can attempt to validate the clinical code sets and the reporting of the conditions. Here is a log for this code set. The prevalence range `64.46% - 67.00%` suggests that this code set is likely well defined.

Date	Practice system	Population	Patients from ID	Patient from code
2021-10-13	EMIS	26929848	1741082 (66.21%)	1741077 (66.21%)
2021-10-13	TPP	211812	137567 (64.95%)	137567 (64.95%)
2021-10-13	Vision	338205	208958 (61.79%)	208958 (61.79%)
LINK: https://github.com/rw251/.../tests/diastolic-blood-pressure/1				

HbA1c

A patient's HbA1c as recorded via clinical code and value. This code set only includes codes that are accompanied by a value (`1003671000000109 - Haemoglobin A1c level`). It does not include codes that indicate a patient's BMI (`165679005 - Haemoglobin A1c (HbA1c) less than 7%`) without giving the actual value.

NB: This code set is intended to indicate a patient's HbA1c. If you need to know whether a HbA1c was recorded then please use v1 of the code set.

Prevalence log

By examining the prevalence of codes (number of patients with the code in their record) broken down by clinical system, we can attempt to validate the clinical code sets and the reporting of the conditions. Here is a log for this code set. The prevalence range `44.93% - 50.88%` suggests that this code set is likely well defined.

Date	Practice system	Population	Patients from ID	Patient from code
2021-05-07	EMIS	2605681	1170688 (44.93%)	1170688 (44.93%)
2021-05-07	TPP	210817	98972 (46.95%)	98972 (46.95%)
2021-05-07	Vision	334632	170245 (50.88%)	170245 (50.88%)

LINK: https://github.com/rw251/.../tests/hba1c/2

Cholesterol

A patient's total cholesterol as recorded via clinical code and value. This code set only includes codes that are accompanied by a value (`44P.. Serum cholesterol`). It does not include codes that indicate a patient's BMI (`44P3. - Serum cholesterol raised`) without giving the actual value.

NB: This code set is intended to indicate a patient's total cholesterol. If you need to know whether a cholesterol was recorded then please use v1 of the code set.

Prevalence log

By examining the prevalence of codes (number of patients with the code in their record) broken down by clinical system, we can attempt to validate the clinical code sets and the reporting of the conditions. Here is a log for this code set. The prevalence range `43.99% - 49.34%` suggests that this code set is likely well defined.

Date	Practice system	Population	Patients from ID	Patient from code
2021-05-11	EMIS	2606497	1146925 (44.00%)	1146651 (43.99%)
2021-05-11	TPP	210810	98627 (46.78%)	98627 (46.78%)
2021-05-11	Vision	334784	165186 (49.34%)	165186 (49.34%)

LINK: https://github.com/rw251/.../tests/cholesterol/2

LDL Cholesterol

A patient's LDL cholesterol as recorded via clinical code and value. This code set only includes codes that are accompanied by a value (`44P6.00 - Serum LDL cholesterol level`).

Prevalence log

By examining the prevalence of codes (number of patients with the code in their record) broken down by clinical system, we can attempt to validate the clinical code sets and the reporting of the conditions. Here is a log for this code set. The prevalence range `41.22% - 45.54%` suggests that this code set is likely well defined.

Date	Practice system	Population	Patients from ID	Patient from code
2021-10-13	EMIS	26929848	1102872 (41.94%)	1102872 (41.94%)
2021-10-13	TPP	211812	91673 (43.28%)	91673 (43.28%)
2021-10-13	Vision	338205	154055 (45.55%)	154055 (45.55%)

LINK: https://github.com/rw251/.../tests/ldl-cholesterol/1

HDL Cholesterol

A patient's HDL cholesterol as recorded via clinical code and value. This code set only includes codes that are accompanied by a value (`44P5.00 - Serum HDL cholesterol level`).

Prevalence log

By examining the prevalence of codes (number of patients with the code in their record) broken down by clinical system, we can attempt to validate the clinical code sets and the reporting of the conditions. Here is a log for this code set. The prevalence range `43.66% - 48.97%` suggests that this code set is likely well defined.

Date	Practice system	Population	Patients from ID	Patient from code
2021-10-13	EMIS	26929848	1168326 (44.42%)	1168326 (44.42%)
2021-10-13	TPP	211812	100823 (47.60%)	100823 (47.60%)
2021-10-13	Vision	338205	165935 (49.06%)	165935 (49.06%)

LINK: https://github.com/rw251/.../tests/hdl-cholesterol/1

Triglyceride (level)

A patient's triglyceride level as recorded via clinical code and value. This code set only includes codes that are accompanied by a value (`44Q..00 - Serum triglycerides`). It does not include codes that indicate a patient's creatinine (`44Q3.00 - Serum triglycerides raised') without giving the actual value.

Codes taken from: https://www.medrxiv.org/content/medrxiv/suppl/2020/05/19/2020.05.14.20101626.DC1/2020.05.14.20101626-1.pdf

Prevalence log

By examining the prevalence of codes (number of patients with the code in their record) broken down by clinical system, we can attempt to validate the clinical code sets and the

reporting of the conditions. Here is a log for this code set. The prevalence range `43.18% - 48.34%` suggests that this code set is likely well defined.

Date	Practice system	Population	Patients from ID	Patient from code
2021-10-13	EMIS	26929848	1135572 (43.18%)	1135572 (43.18%)
2021-10-13	TPP	211812	99188 (46.82%)	99188 (46.82%)
2021-10-13	Vision	338205	163502 (48.34%)	163502 (48.34%)

LINK: https://github.com/rw251/.../tests/triglycerides/1

Glomerular filtration rate (GFR)

Any code that gives the value of a patient's GFR (or estimated GFR - EGFR).

Prevalence log

By examining the prevalence of codes (number of patients with the code in their record) broken down by clinical system, we can attempt to validate the clinical code sets and the reporting of the conditions. Here is a log for this code set. The prevalence range `52.43% - 57.73%` suggests that this code set is likely well defined.

Date	Practice system	Population	Patients from ID	Patient from code
2021-06-10	EMIS	2610073	1369349 (52.46%)	1368468 (52.43%)
2021-06-10	TPP	211034	121897 (57.76%)	121835 (57.73%)
2021-06-10	Vision	335344	184635 (55.06%)	184523 (55.02%)

LINK: https://github.com/rw251/.../tests/egfr/1

Height

A patient's height as recorded via clinical code and value. This code set only includes codes that are accompanied by a value (`229.. - O/E - Height`).

Codes taken from https://www.medrxiv.org/content/medrxiv/suppl/2020/05/19/2020.05.14.20101626.DC1/2020.05.14.20101626-1.pdf

Prevalence log

By examining the prevalence of codes (number of patients with the code in their record) broken down by clinical system, we can attempt to validate the clinical code sets and the

reporting of the conditions. Here is a log for this code set. The prevalence range `66.10% - 72.59%` suggests that this code set is well defined.

Date	Practice system	Population	Patients from ID	Patient from code
2021-10-13	EMIS	26929848	1885015 (71.68%)	1884110 (71.64%)
2021-10-13	TPP	211812	140013 (66.10%)	140013 (66.10%)
2021-10-13	Vision	338205	245440 (72.59%)	245440 (72.57%)
LINK: https://github.com/rw251/.../patient/height/1				

Weight

A patient's weight as recorded via clinical code and value. This code set only includes codes that are accompanied by a value.

Codes taken from https://www.medrxiv.org/content/medrxiv/suppl/2020/05/19/2020.05.14.20101626.DC1/2020.05.14.20101626-1.pdf

Prevalence log

By examining the prevalence of codes (number of patients with the code in their record) broken down by clinical system, we can attempt to validate the clinical code sets and the reporting of the conditions. Here is a log for this code set. The prevalence range `73.09% - 79.68%` suggests that this code set is well defined.

Date	Practice system	Population	Patients from ID	Patient from code
2021-10-13	EMIS	26929848	2054449 (78.12%)	2053717 (78.09%)
2021-10-13	TPP	211812	154813 (73.09%)	154813 (73.09%)
2021-10-13	Vision	338205	269496 (79.68%)	269496 (79.68%)
LINK: https://github.com/rw251/.../patient/weight/1				

Body Mass Index (BMI)

A patient's BMI as recorded via clinical code and value. This code set only includes codes that are accompanied by a value (`22K.. - Body Mass Index`). It does not include codes that indicate a patient's BMI (`22K6. - Body mass index less than 20`) without giving the actual value.

NB: This code set is intended to indicate a patient's BMI. If you need to know whether a BMI was recorded then please use v1 of the code set.

Prevalence log

By examining the prevalence of codes (number of patients with the code in their record) broken down by clinical system, we can attempt to validate the clinical code sets and the reporting of the conditions. Here is a log for this code set. The prevalence range `63.96% - 79.69%` suggests that this code set is perhaps not well defined. However, as EMIS (80% of practices) and TPP (10% of practices) are close, it could simply be down to Vision automatically recording BMIs and therefore increasing the prevalence there.

UPDATE By looking at the prevalence of patients with a BMI code that also has a non-zero value the range becomes `62.48% - 64.93%` which suggests that this code set is well defined.

Date	Practice system	Population	Patients from ID	Patient from code
2021-05-07	EMIS	2605681	1709250 (65.60%)	1709224 (65.60%)
2021-05-07	TPP	210817	134841 (63.96%)	134835 (63.96%)
2021-05-07	Vision	334632	266612 (79.67%)	266612 (79.67%)
2021-05-11	EMIS	2606497	1692442 (64.93%)	1692422 (64.93%)
2021-05-11	TPP	210810	134652 (63.87%)	134646 (63.87%)
2021-05-11	Vision	334784	209175 (62.48%)	209175 (62.48%)

LINK: https://github.com/rw251/.../patient/bmi/2

Smoking status current

Any code suggestive that a patient is a current smoker.

LINK: https://github.com/rw251/.../patient/smoking-status-current/1

Smoking status currently not

Any code suggestive that a patient is currently a non-smoker. This is different to the "never smoked" code set.

LINK: https://github.com/rw251/.../patient/smoking-status-currently-not/1

Smoking status ex

Any code suggestive that a patient is an ex-smoker.

LINK: https://github.com/rw251/.../patient/smoking-status-ex/1

LINK: https://github.com/rw251/.../patient/smoking-status-ex-trivial/1

Smoking status never

Any code suggestive that a patient has never smoked. This is different to the "currently not" code set.

LINK: https://github.com/rw251/.../patient/smoking-status-never/1

LINK: https://github.com/rw251/.../patient/smoking-status-passive/1

LINK: https://github.com/rw251/.../patient/smoking-status-trivial/1

Prevalence log

By examining the prevalence of codes (number of patients with the code in their record) broken down by clinical system, we can attempt to validate the clinical code sets and the reporting of the conditions. Here is a log for this code set.

The discrepancy between the patients counted when using the IDs vs using the clinical codes is due to these being new codes which haven't all filtered through to the main Graphnet dictionary. The prevalence range `1.19% - 26.55%` as of 11th March 2021 is too wide. However the prevalence figure of 26.55% from EMIS is close to public data and is likely ok.

UPDATE - 25th March 2021 Missing Read and CTV3 codes were added to the vaccination list and now the range of `26.91% - 32.96%` seems reasonable. It should be noted that there is an approx 2 week lag between events occurring and them being entered in the record.

UPDATE - 12th April 2021, latest prevalence figures.

UPDATE - 18th March 2022 There are now new codes for things like 3rd/4th/booster dose of vaccine. The latest prevalence shows `65.0% - 66.3%` have at least one vaccine code in the GP_Events table, and `88.2% - 93.6%` have at least one code for the vaccine in the GP_Medications table.

MED

Date	Practice system	Population	Patients from ID	Patient from code
2021-05-12	EMIS	2606497	0 (0%)	379577(14.56%)
2021-05-12	TPP	210810	0 (0%)	1637(0.78%)
2021-05-12	Vision	334784	0 (0%)	93(0.03%)
2022-03-18	EMIS	2658131	1750506 (65.9%)	1763420(66.3%)
2022-03-18	TPP	212662	8207 (3.9%)	138285(65.0%)
2022-03-18	Vision	341594	122060 (35.7%)	225844(66.1%)

EVENT

Date	Practice system	Population	Patients from ID	Patient from code
2021-05-12	EMIS	2606497	4446 (0.17%)	1101577 (42.26%)
2021-05-12	TPP	210810	7 (0.00%)	87841 (41.66%)
2021-05-12	Vision	334784	1 (0.00%)	142724 (42.63%)
2022-03-18	EMIS	2658131	2486786 (93.6%)	1676951 (63.1%)
2022-03-18	TPP	212662	187463 (88.2%)	7314 (3.44%)
2022-03-18	Vision	341594	312617 (91.5%)	62512 (18.3%)

LINK: https://github.com/rw251/.../procedures/covid-vaccination/1

Clinical code sets

All code sets required for this analysis are available here: https://github.com/rw251/.../046 - Gupta/clinical-code-sets.csv. Individual lists for each concept can also be found by using the links above.