CPRD COVID-19 Synthetic Data Specification

Version 4.0

Date: 23 November 2021

Documentation Control Sheet

It may be necessary to issue amendments or clarifications to parts of this document. This form must be updated whenever changes are made and should be filled inside the front cover of the new or amended document.

Version	Summary of Change	Prepared By	Date	Reviewed By	Date
1.0		J. E. de Benedetti	18/06/2020	Puja Myles Jessie Oyinlola	25/06/2020
2.0	Coverage period and counts updated	J. E. de Benedetti	13/11/2020		
3.0	Coverage period and counts updated; Addition of ‘Is Covid test Positive’ Boolean Variable	J. E. de Benedetti	15/04/2021	Puja Myles	16/04/2021
4.0	Coverage period and counts updated	Barbara Draghi	23/11/2021	Puja Myles	25/11/2021

About the Dataset

This wholly synthetic dataset is based on real anonymised primary care patient data extracted from the CPRD Aurum database. Researchers will not be able to access the real anonymised patient data extract which were used as the basis for the synthetic dataset generation to preserve patient privacy.

The synthetic dataset focuses on patients presenting to primary care with symptoms indicative of COVID-19 (confirmed/suspected Covid-19) and control patients with negative COVID-19 test results. The dataset includes data on sociodemographic and clinical risk factors. The ‘ground truth’ CPRD Aurum data extract used as the basis for generating this synthetic dataset included data from 03/12/2019 till 22/11/2021 on patients with suspected, confirmed or negative COVID-19 as ascertained from the primary care record. The ground truth data extract was subject to data pre-processing and as such, the synthetic dataset based on this, does not reflect the structure of the source CPRD Aurum database.

The development of this synthetic dataset was funded by NHS X using the synthetic data generation and evaluation framework developed by CPRD under a grant from the Regulators’ Pioneer Fund launched by The Department for Business, Energy and Industrial Strategy (BEIS) and managed by Innovate UK. The methodology used to generate and evaluate this synthetic dataset is outlined in Wang et al. 2019.

Note: There may be some COVID diagnoses recorded after date of death, which reflects the observations in the ground truth data.

Count

Total patients: 4,173,000

Patients with negative result test: 3,436,379
Patients with confirmed/suspected COVID-19: 736,621

Data Format

1. The dataset for the Synthetic data (file xxx_yyyCovidSynData.csv) is in a comma separated values (csv) format. It is separated by the character ‘,’. In the title of the file, the “xxx” refers to the version number and it can be between 001 and 999. The “yyy” refers to the number of versions of the ground truth file.

2. The **rural/urban** file (urbanClasification.txt) contains the description for the values of the column rurban in the dataset. The format is csv with the character ‘,’ as the separator field. This file can be linked back to the Synthetic Covid19 data via rurban id.

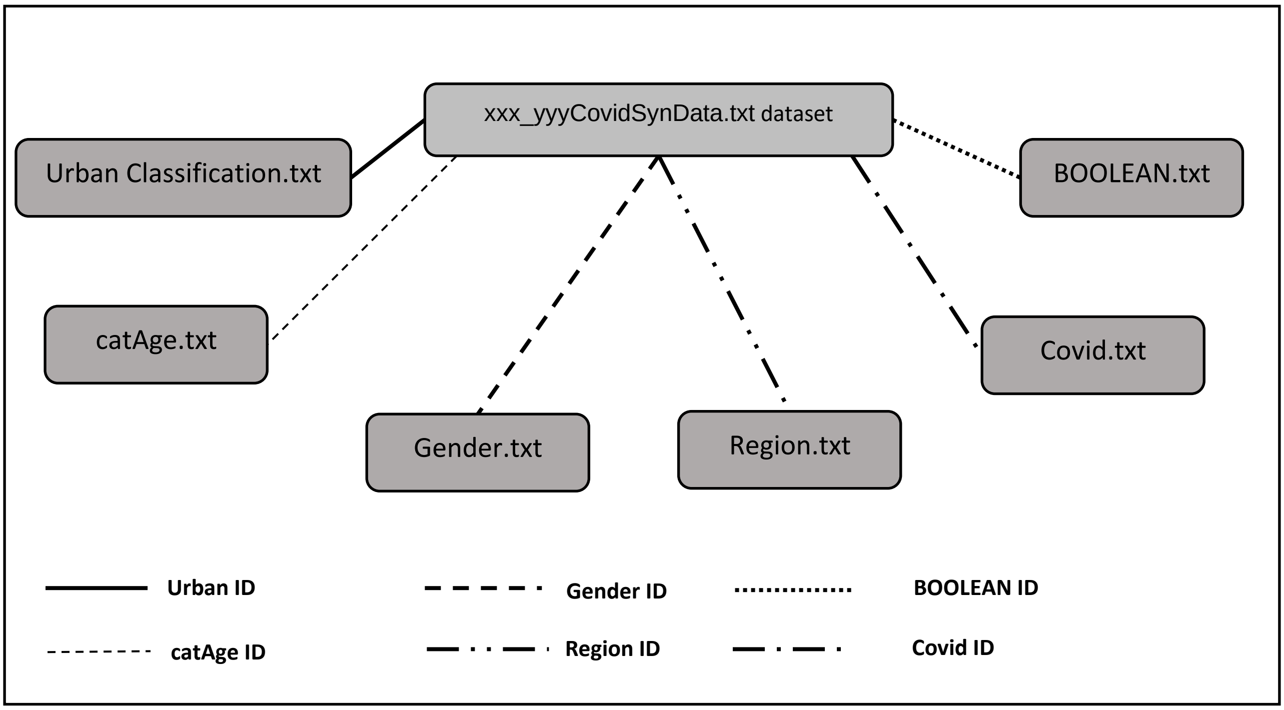
3. The **region** file (Region.txt) contains the description for the values of the column region in the dataset. The format is csv with the character ‘,’ as the separator field. This file can be linked back to the Synthetic Covid19 data via region id.

4. The **gender** file (Gender.txt) contains the description for the values of the column gender in the dataset. The format is csv with the character ‘,’ as the separator field. This file can be linked back to the Synthetic Covid19 data via gender id.

5. The file **catAge**.txt contains the description for the values of the column catAge in the dataset. The format is csv with the character ‘,’ as the separator field. This file can be linked back to the Synthetic Covid19 data via catAge id.

6. For all the **binary columns** on the dataset, the file BOOLEAN.txt contains the description of the values. The format is csv with the character ‘,’ as the separator field. This file can be linked back to the Synthetic Covid19 data via boolean id.

7. The Covid file (Covid.txt) contains the description for the values of the column covid_diagnosis in the dataset. The format is csv with the character ‘,’ as the separator field. This file can be linked back to the Synthetic Covid19 data via covid id.



Field Descriptions

Full descriptions of the fields in each data file are provided in the tables below. The look-up column lists the look-up files (presented later in this document) with further information on decoding numerical values.

Column name	*Description*	*Look-up*	*Type*	*Format*
region	Region in the UK	Region	Categorical Nominal	Integer Numbers
gender	Gender of the patient	Gender	Categorical Nominal	Integer Numbers
covid_dt	Date of the latest COVID event recorded.	n/a	Numerical	Date: ddmmyy
age	Age of the patient based on birthyear	n/a	Numerical	Integer Numbers
agecat	Age category	catAge	Categorical Ordinal	Integer Number
imd_5	Index of multiple deprivation quintile- at practice level (1=LEAST deprived; 5= MOST deprived)	n/a	Categorical Ordinal	Integer Numbers
dateDeath	Date of estimated death.	n/a	Numerical	Date: ddmmyy
Death	Confirms patient death.	BOOLEAN	Categorical Binary	Integer Number
rurban	General Practice location (Rural /Urban)	Urban classification	Numerical	Integer Number
CVDrx	Medication used to treat cardiovascular disease: A prescription of an angiotensin-converting enzyme inhibitor (ACEi) or Angiotensin II receptor blocker (ARB)	BOOLEAN	Categorical Binary	Integer Number
Resprx	Medication used to treat asthma or Chronic Obstructive Pulmonary Disease (COPD): A prescription of an inhaled corticosteroids (ICS) or short-acting beta-agonists (SABA) or long-acting beta-agonists (LABA) or long-acting muscarinic antagonists (LAMA) or short-acting muscarinic antagonists (SAMA) or ICS+LABA orICS+SABA or LAMA+LABA or Theophylline.	BOOLEAN	Categorical Binary	Integer Number
CVD_Resp_rx	Medication used to treat either cardiovascular disease, asthma or COPD: A prescription of an ACEi or ARB or ICS or SABA or LABA or LAMA or SAMA or ICS+LABA or ICS+SABA orLAMA+LABA or Theophylline	BOOLEAN	Categorical Binary	Integer Number

Column name	*Description*	*Look-up*	*Type*	*Format*
ARB_rx	Medication: A prescription of an ARB	BOOLEAN	Categorical Binary	Integer Number
ACEi_rx	Medication: A prescription of an ACEi	BOOLEAN	Categorical Binary	Integer Number
SAMA_rx	Medication: A prescription of a SAMA	BOOLEAN	Categorical Binary	Integer Number
LAMA_rx	Medication: A prescription of a LAMA	BOOLEAN	Categorical Binary	Integer Number
SABA_rx	Medication: A prescription of a SABA	BOOLEAN	Categorical Binary	Integer Number
LABA_rx	Medication: A prescription of a LABA	BOOLEAN	Categorical Binary	Integer Number
ICS_rx	Medication: A prescription of an ICS	BOOLEAN	Categorical Binary	Integer Number
AminoTheophy_rx	Medication: A prescription of Theophylline	BOOLEAN	Categorical Binary	Integer Number
LAMALABA_rx	Medication: A prescription of a LAMA+LABA	BOOLEAN	Categorical Binary	Integer Number
ICSLABA_rx	Medication: A prescription of an ICS+LABA	BOOLEAN	Categorical Binary	Integer Number
Inmuno_rx	Medication: A prescription for an immunosuppressant (either a corticosteroid or janus kinase inhibitor ormammalian target of rapamycin (mTOR) inhibitor orinosine monophosphate dehydrogenase (IMDH) inhibitor or biologic or monoclonal antibodies or sphingosine-1-phosphate receptor modulator)	BOOLEAN	Categorical Binary	Integer Number
Tamiflu_rx	Medication: A prescription of Tamiflu (Oseltamivir)	BOOLEAN	Categorical Binary	Integer Number
Chloro_Hydroxychloro_rx	Medication: A prescription for Hydroxychloroquine, Chloroquine or Quinine	BOOLEAN	Categorical Binary	Integer Number
fatmyalgia	Presenting symptom: fatigue/myalgia	BOOLEAN	Categorical Binary	Integer Number
fever	Presenting symptom: fever	BOOLEAN	Categorical Binary	Integer Number
cough	Presenting symptom: cough	BOOLEAN	Categorical Binary	Integer Number
MI	Comorbidity: myocardial infarction	BOOLEAN	Categorical Binary	Integer Number

Column name	*Description*	*Look-up*	*Type*	*Format*
LiverDis	Comorbidity: liver disease	BOOLEAN	Categorical Binary	Integer Number
StrokeTIA	Comorbidity: stroke/transient ischaemic attack	BOOLEAN	Categorical Binary	Integer Number
RheumatoidArthritis	Comorbidity: rheumatoid arthritis	BOOLEAN	Categorical Binary	Integer Number
PAD	Comorbidity: peripheral artery disease (PAD)	BOOLEAN	Categorical Binary	Integer Number
LearningDisability	Comorbidity: learning disability	BOOLEAN	Categorical Binary	Integer Number
Hypertension	Comorbidity: hypertension	BOOLEAN	Categorical Binary	Integer Number
HeartFailure	Comorbidity: heart failure	BOOLEAN	Categorical Binary	Integer Number
Epilepsy	Comorbidity: epilepsy	BOOLEAN	Categorical Binary	Integer Number
Diabetes	Comorbidity: type 1 and type 2 diabetes	BOOLEAN	Categorical Binary	Integer Number
MentalHealth	Comorbidity: mental health condition	BOOLEAN	v Categorical Binary	Integer Number
Depression	Comorbidity: depression	BOOLEAN	Categorical Binary	Integer Number
Dementia	Comorbidity: dementia	BOOLEAN	Categorical Binary	Integer Number
COPD	Comorbidity: COPD	BOOLEAN	Categorical Binary	Integer Number
CKD	Comorbidity: chronic kidney disease	BOOLEAN	Categorical Binary	Integer Number
Cancer	Comorbidity: cancer	BOOLEAN	Categorical Binary	Integer Number
Asthma	Comorbidity: asthma	BOOLEAN	Categorical Binary	Integer Number
AF	Comorbidity: atrial fibrillation	BOOLEAN	Categorical Binary	Integer Number
PalliativeCare	Receiving palliative care	BOOLEAN	Categorical Binary	Integer Number
covid_diagnosis	Suspected/confirmed	Covid	Categorical Binary	Integer Number

Column name	*Description*	*Look-up*	*Type*	*Format*
isPositive	If Negative Covid test result, then variable is 0 (False)	BOOLEAN	Categorical Binary	Integer Number

Look-up files

Region

Id	*Description*
1	North East
2	North West
3	Yorkshire And Humber
4	East Midlands
5	West Midlands
6	East of England
7	South West
8	South Central
9	London
10	South East Coast
11	Northern Ireland

Gender

Id	*Description*
1	Male
2	Female

Urban Classification

Id	*Description*
1	Living in urban area
2	Living in rural area

catAge

Id	*Description*
1	<20 years old.
2	20-44 years old.
3	45-69 years old.
4	70+ years old.

BOOLEAN

Id	*Description*
0	No
1	Yes

Covid

Id	*Description*
0	Suspected
1	Confirmed