CPRD Synthetic Cardiovascular Disease Data Specification

Version 1.0

Date: 1 June 2020

Author: Namir Oues

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		Namir Oues	25/03/2020	Zhenchen Wang	01/06/2020

About the Dataset

This wholly synthetic dataset is based on real anonymised primary care patient data extracted from the CPRD Aurum database and focuses on cardiovascular disease risk factors. 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 ground truth data extract was subject to data pre-processing and as such, the synthetic dataset, which is based on this, does not reflect the structure of the source CPRD Aurum database.

This synthetic dataset was developed as part of a project funded by 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.

Count

Total Patients: 499,344

Data Format

The data are available to researchers as 5 files in text format as listed below:

1. The **Synthetic Cardiovascular Disease** file (SyntheticCVD.txt) that contains the medical information for all the patients.

2. The **Ethnicity** file (Ethnicity.txt) is a lookup that contains all different ethnicities. This file can be linked back to the Synthetic data via ethnicity id.

3. The **Smoking** file (Smoking.txt) is a lookup that contains information on the smoking status. This file can be linked back to the synthetic data via smoking id.

4. The **Gender** file (Gender.txt) is a lookup that contains all different genders. This file can be linked back to the synthetic data via gender id.

5. The **Boolean Variables** file (BooleanVariables.txt) is a lookup that contains two variables for positive or negative. This file can be linked back to the synthetic data via Boolean id.

Field Descriptions

Full descriptions of the fields in each data file are provided in the tables below. The mapping column lists lookup files with further information on decoding numerical values.

Column name	*Field name*	*Description*	*Mapping*	*Type*
Stroke or heart attack	strokeha	Stroke indicator for the patient	Lookup: BooleanVariables.txt	Categorical binary
Age	age	Patient's age	n/a	Numeric
Atrial fibrillation	af	Atrial fibrillation indicator for patient	Lookup: BooleanVariables.txt	Categorical binary
Atypical antipsychotic medication	atyantip	Indicates if patient is on atypical antipsychotic medication	Lookup: BooleanVariables.txt	Categorical binary
Steroid tablets	steroid	Indicates if patient is on regular steroid tablets	Lookup: BooleanVariables.txt	Categorical binary
Erectile dysfunction	impot	A gender specific value that indicates if a male patient has erectile dysfunction	Lookup: BooleanVariables.txt	Categorical binary
Migraines	migr	Migraines indicator for patient	Lookup: BooleanVariables.txt	Categorical binary
Rheumatoid arthritis	ra	Indicator of rheumatoid arthritis for the patient	Lookup: BooleanVariables.txt	Categorical binary
Chronic kidney disease	ckidney	Chronic kidney disease indicator	Lookup: BooleanVariables.txt	Categorical binary
Severe mental illness	semi	Severe mental illness indicator for patient (this includes schizophrenia, bipolar disorder and moderate/severe depression)	Lookup: BooleanVariables.txt	Categorical binary
Systemic lupus erythematosus	sle	Systemic lupus erythematosus indicator for patient	Lookup: BooleanVariables.txt	Categorical binary
Blood pressure treatment	treathyp	Indicates if patient is on blood pressure treatment	Lookup: BooleanVariables.txt	Categorical binary
Diabetes type 1	type1	Type1 diabetes indicator for patient	Lookup: BooleanVariables.txt	Categorical binary
Diabetes type 2	type2	Type2 diabetes indicator for patient	Lookup: BooleanVariables.txt	Categorical binary
Body Mass Index	bmi	Body mass index for patient	n/a	Numeric
Ethnicity	ethr	Value that indicates the ethnicity of the patient	Lookup: Ethnicity.txt	Categorical nominal
Cholesterol ratio	choleratio	Value for cholesterol ratio for patient	n/a	Numeric
Systolic blood pressure	sbp	Value for systolic blood pressure for patient (measured in mmHg)	n/a	Numeric

Systolic blood pressure standard deviation	sbps	Standard deviation of at least two most recent systolic blood pressure readings for a patient (measured in mmHg):	n/a	Numeric
Smoking	smoking	Value that indicates the smoking status of a patient	Lookup: SmokingStatus.txt	Categorical nominal
Gender	gender	Patient's gender	Lookup: Gender.txt	Categorical nominal
Region	region	Value that indicates where in the UK the patient lives	n/a	Categorical nominal

2. Ethnicity

Id	*Description*
1	White or not stated
2	Indian
3	Pakistani
4	Bangladeshi
5	Other Asian
6	Black Caribbean
7	Black African
8	Chinese
9	Other ethnic group

3. Smoking

Id	*Description*
0	Non-smoker
1	Ex-smoker
2	Light smoker
3	Moderate smoker
4	Heavy smoker

4. Gender

Id	*Description*
M	Male
F	Female
I	Intersex
U	Unknown

5. Boolean variables

Id	*Description*
0	No
1	Yes