Medicines & Healthcare products Regulatory Agency

CPRD | UK data driving real-world evidence

Release Notes: CPRD GOLD Sample Dataset April 2023

Summary

The CPRD GOLD synthetic dataset is a medium-fidelity synthetic dataset that resembles[1] the real world CPRD GOLD with respect to the data types, data values, data formats, data structure and table relationships. This synthetic dataset can be used for multiple purposes including:

1. as a sample dataset to understand the structure and utility of the anonymised CPRD GOLD database
2. as a data management teaching/training resource
3. to develop/validate/test analytics tools for use with CPRD GOLD data
4. to improve bespoke CPRD GOLD application interfaces/algorithms, e.g. a bespoke cohort selection tool, or
5. to develop machine learning workflows that can be applied to anonymised CPRD GOLD data.

The table below presents summary statistics for the synthetic CPRD GOLD dataset[7].

Metric	Coverage
Total number of acceptable patients[2] (including transferred out and deceased patients):	103,093
Current acceptable patients (i.e. registered at currently contributing practices, excluding transferred out and deceased patients):	91,780
Percentage UK population coverage[3] (current patients only):	91,780 of 67,026,300 (0.13%)

Available follow-up time in years since 1st January 1995[4] (all patients including transferred out and deceased):
Mean (Standard Deviation):
Median (25th and 75th Percentile):
	
	16.81 (15.93)
	13.25 (3.81 – 25.07)

Available follow-up time in years since 1st January 1995 (current patients only):	23.01 (14.54)

[1] Primary/Foreign keys of tables may not be in the same format as real CPRD GOLD data specs.

[2] Permanent registrations only. Over 98% of permanent registrations are deemed to have 'acceptable' (or research quality) data based on CPRD metrics.

[3] Based on latest UK population estimates from the Office of National Statistics.

[4] Follow-up time stated here does not incorporate the up-to-standard (UTS) date and the database includes records pre-dating the 1st of January 1995.

Mean (Standard Deviation):
Median (25th and 75th Percentile):
	20.61 (13.20 – 29.71)

Total number of practices (current and historic) included in the database:	14
Currently contributing practices[5]:	14
Percentage coverage of UK general practices (currently contributing practices only):	14 of 8,104 (0.17%)
Regional distribution of currently contributing practices[6]	
North East:	2
North West:	3
Yorkshire And The Humber:	0
East Midlands:	0
West Midlands:	1
East of England:	3
South West:	1
South Central:	0
London:	1
South East Coast:	1
Northern Ireland:	0
Scotland:	1
Wales:	1

DOI

Please cite in any publications using this version: https://doi.org/10.48329/y7q8-gr42

CPRD GOLD Data Specification

CPRD GOLD Sample Dataset Specification is similar to CPRD GOLD Specification which is available on the CPRD website: https://www.cprd.com/primary-care

[5] A practice that has contributed data to CPRD within 60 days of the database build being created.

[6] Expressed as number of practices contributing to CPRD GOLD.

[7] Numbers are based on the CPRD GOLD April 2023 release https://doi.org/10.48329/r4kf-ax47