CINECA synthetic cohort Europe CH SIB

Data access guide, v1

Nona Naderi, Douglas Teodoro
May 2021

Dataset description

The “CINECA synthetic cohort EUROPE CH SIB” dataset consists of 6733 synthetic samples with phenotypic and genotypic information. The synthetic phenotypic data were created from the CoLaus and PsyColaus cohort and the synthetic genetic data from the 1000 Genomes project. CoLaus and PsyColaus cohorts include data from more than 6k Causasian individuals aged 35 to 75 years living in Lausanne, Switzerland. CoLaus focuses on cardiovascular disorders and PsyCoLaus focuses on psychiatric disorders. CoLaus/PsyCoLaus cohort collects demographic, socio-economical, life-style, and clinical information from enrolled patients. While CoLaus/PsyCoLaus contains phenotypic data, the 1000 Genomes data was used to minimise the used data and improve patient’s privacy.

To generate the phenotypic synthetic data, the DataSynthesizer tool was used. This tool is specifically designed for privacy-preserving datasets and enables the generation of randomly and statistically correlated distributions for the synthetic data. To minimise the data used, 20 variables out of 191 available in the original dataset were selected based on their relevance to the CINECA metadata model. The variables encode the following information: *age* (numeric), *gender* (categorical numeric, woman 0, man 1), *birthplace* (categorical string), *residence* (categorical string), *job type* (categorical string), *family and household structure* (categorical numeric, alone 0, couple 1); *tobacco* (categorical string), *alcohol use* (categorical string), and *physical activity* (categorical string); *weight* (numeric), *height* (numeric), *blood pressure* (numeric) and *heart rate* (numeric); and *diagnoses* (free text, string) and *prescriptions* (ATC codes, string). These values were generated using a random distribution. The “subjects” of this generated synthetic data were marked with the term *FAKE.*

The synthetic genetic data contains 100 samples extracted from the 1000 Genome (release 20130502) VCF file, where the variant position is lower than 16070000. The genetic data was then randomly linked to the phenotypic data.

This synthetic dataset has no identifiable data and cannot be used to make any inference about CoLaus and PsyCoLaus cohort data or results.

Acknowledgement:
We thank Loïc Schüpbach for his contribution to set up the data synthesizer tool locally.

CINECA project is funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No 825775 and the Canadian Institute of Health Research under CIHR grant number 404896.

Data access guide

Zenodo