a Data Resources

For Development

All of Us
RESEARCH PROGRAM

For Validation

biobank uk

Electronic Health Record
e.g.,
· conditions
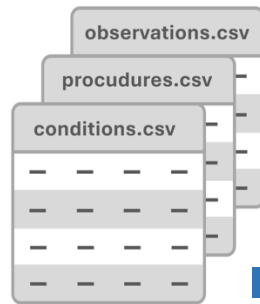· medications

Whole Genome Sequencing

1257 identified independent genome-wide significant SNPs
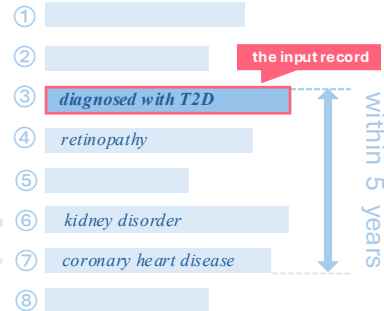
b Data Preprocessing

EHR Data under OMOP CDM

observations.csv
procudures.csv
conditions.csv

Flattening

Per-patient Visit Timeline

①
②
③ diagnosed with T2D — the input record
④ retinopathy
⑤
⑥ kidney disorder
⑦ coronary heart disease
⑧

within 5 years

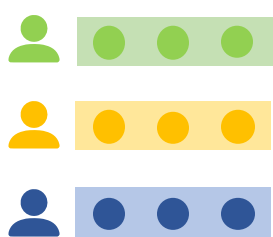1 0 0 1 0 1
multi-hot labeling

c Hypergraph-based Prediction & Subtyping
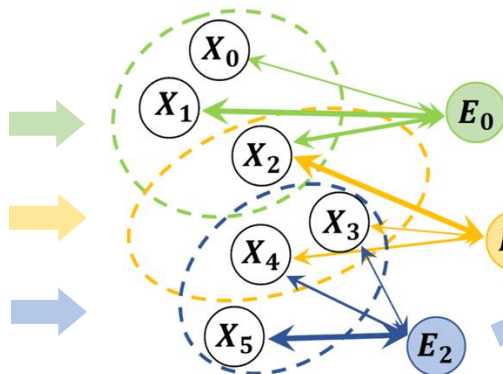
Model Input
initial T2D diagnosis visit for each patient (diabetic codes excluded)
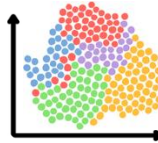
Records

Hypergraph Learning
EHR codes as nodes, patient visits as hyperedges

X_0 X_1 X_2 X_3 X_4 X_5 E_0 E_1 E_2

Self-supervised Clustering

DEC

clusters assignment

Predicting Complications

MLP

predicted labels
0 0 1 1 0 1
1 1 0 1 0 1
1 0 0 1 0 1