Generating Synthetic Electronic Health Records in OMOP using GPT

Chao Pang, Xinzhuo Jiang, Nishanth Parameshwar Pavinkurve, Krishna S. Kalluri, Elise L. Minto, Jason Patterson, Karthik Natarajan

Department of Biomedical Informatics

Columbia University

OHDSI

Observational Health Data Sciences and Informatics

Motivations for synthetic EHR data

Machine Learning

- Prediction research
- External validation

Phenotype algorithm validation

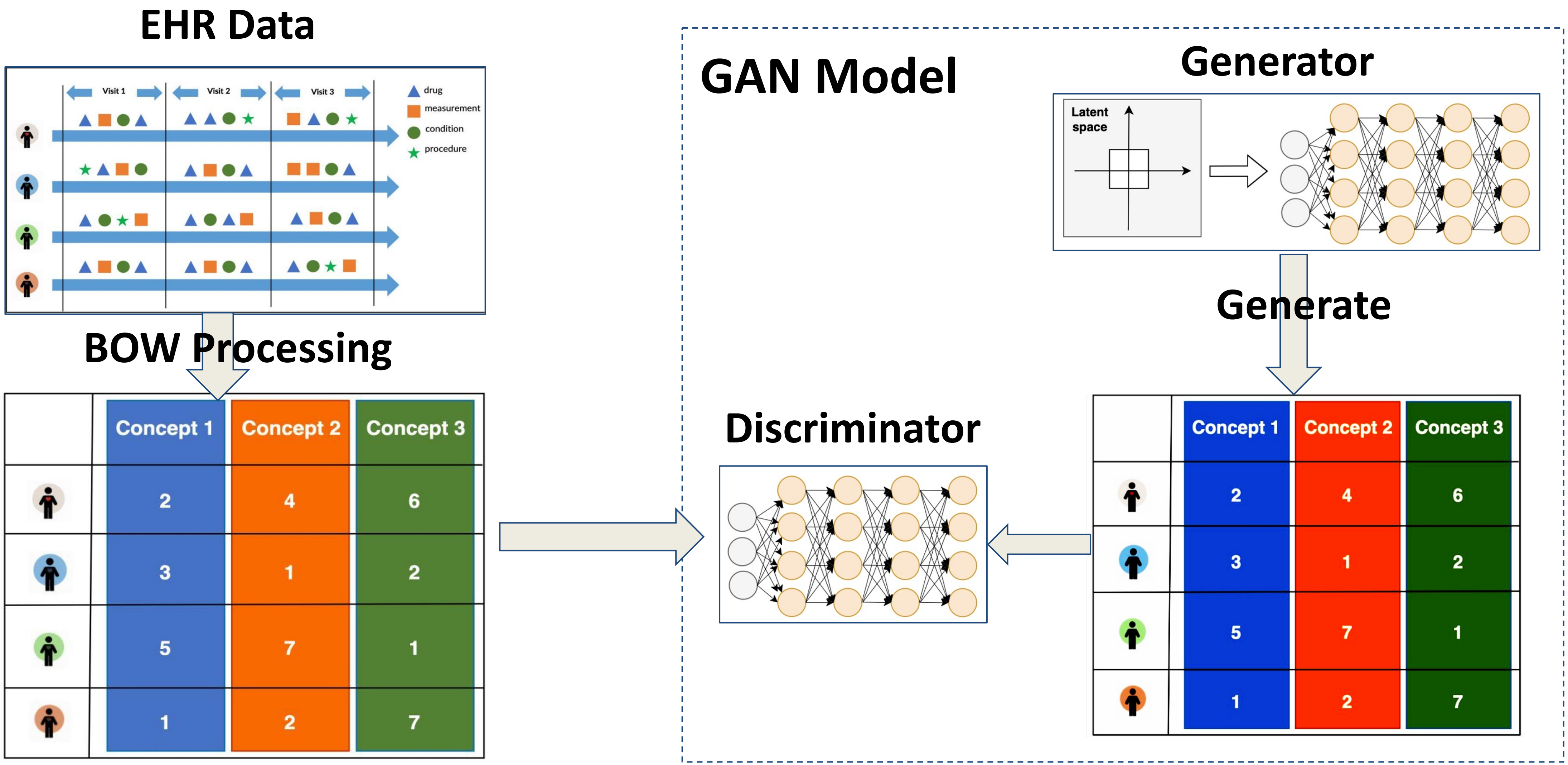
Tool development

Training and education

Fairness and Bias

- Debiasing the source data
- Counterfactual dataset

Common Approach: Bag of Word (BOW) + GAN



JOURNAL ARTICLE

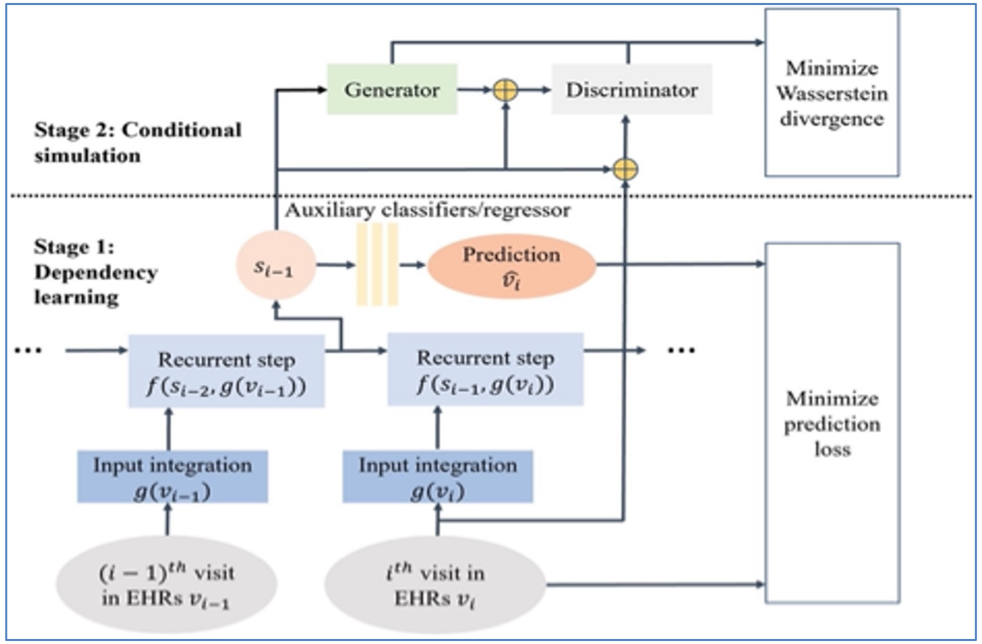
SynTEG: a framework for temporal structured electronic health data simulation FREE

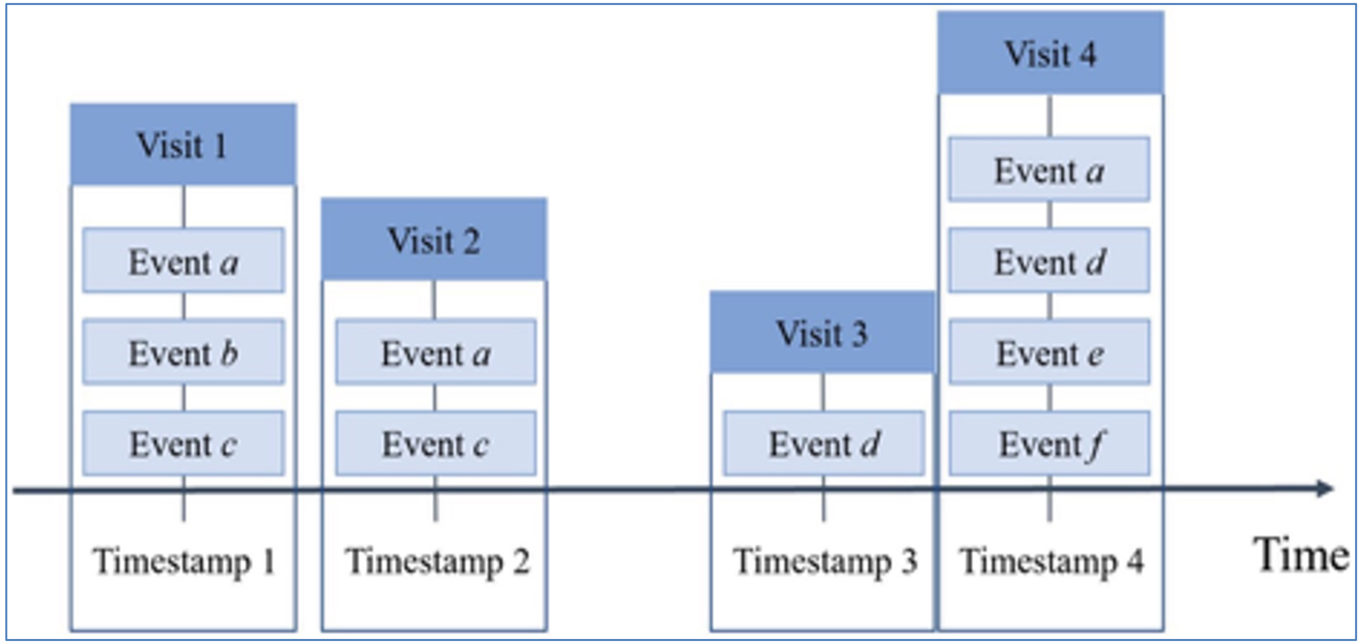
Ziqi Zhang, Chao Yan, Thomas A Lasko, Jimeng Sun, Bradley A Malin

Journal of the American Medical Informatics Association, Volume 28, Issue 3, March 2021, Pages 596–604,
https://doi.org/10.1093/jamia/ocaa262

Published: 23 November 2020 **Article history** ▾

PDF Split View Cite Permissions Share ▾

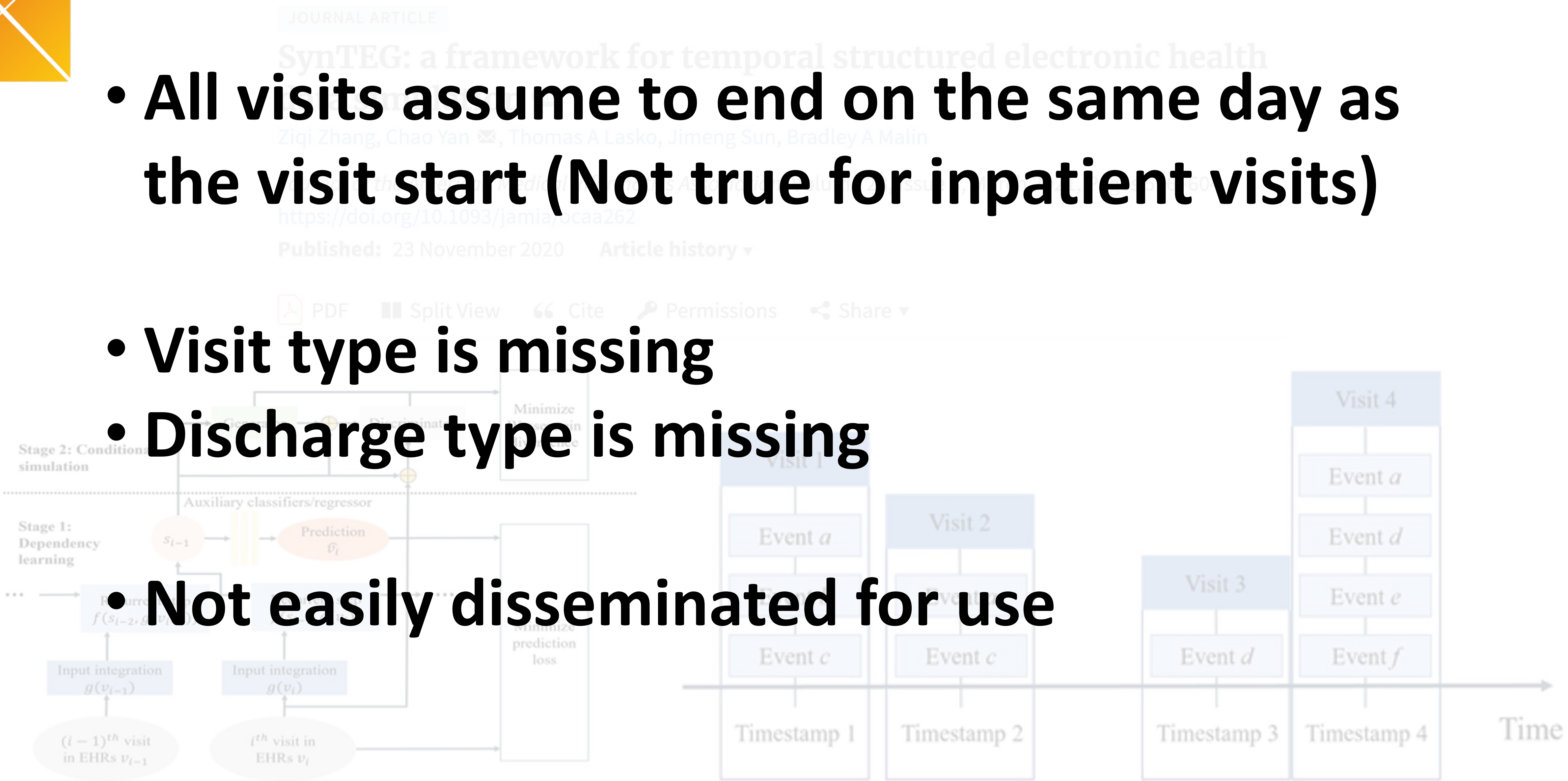




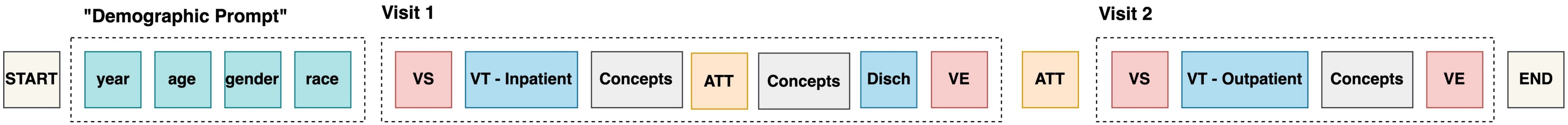
- All visits assume to end on the same day as the visit start (Not true for inpatient visits)

- Visit type is missing
- Discharge type is missing

- Not easily disseminated for use



Patient Representation



year — Year at first visit

age — Age at first visit

gender — Gender

race — Race

VS — Visit Start

VE — Visit End

VT — Visit Type

Disch — Discharge type

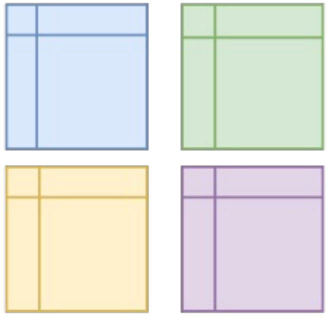
ATT — Artificial Time Token
Day token

Concepts — Condition, Drug
Ingredient, Procedure

CEHR-BERT https://proceedings.mlr.press/v158/pang21a/pang21a.pdf

Proposed Synthetic Data Framework

OMOP

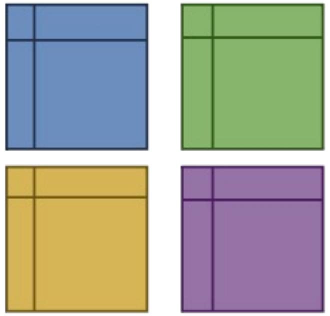


Evaluation Framework

OHDSI

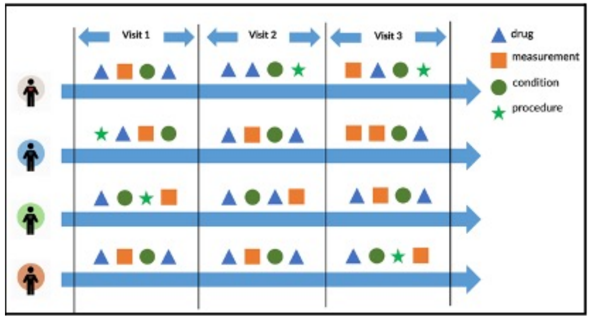
- Concept Prevalence
- Atlas
  - Achilles
  - Cohort Characterization
- Co-occurrence Metrics
- ML Prediction Performance (PLP)
- Data Privacy

Synthetic OMOP

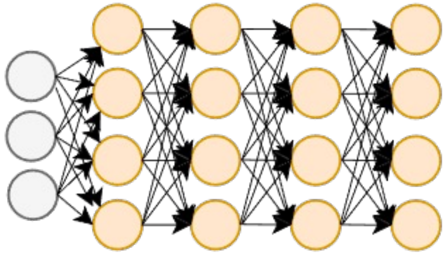


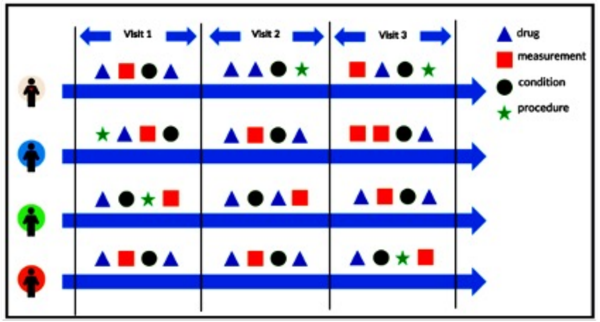
OMOP Converter

Convert OMOP to Patient Representation

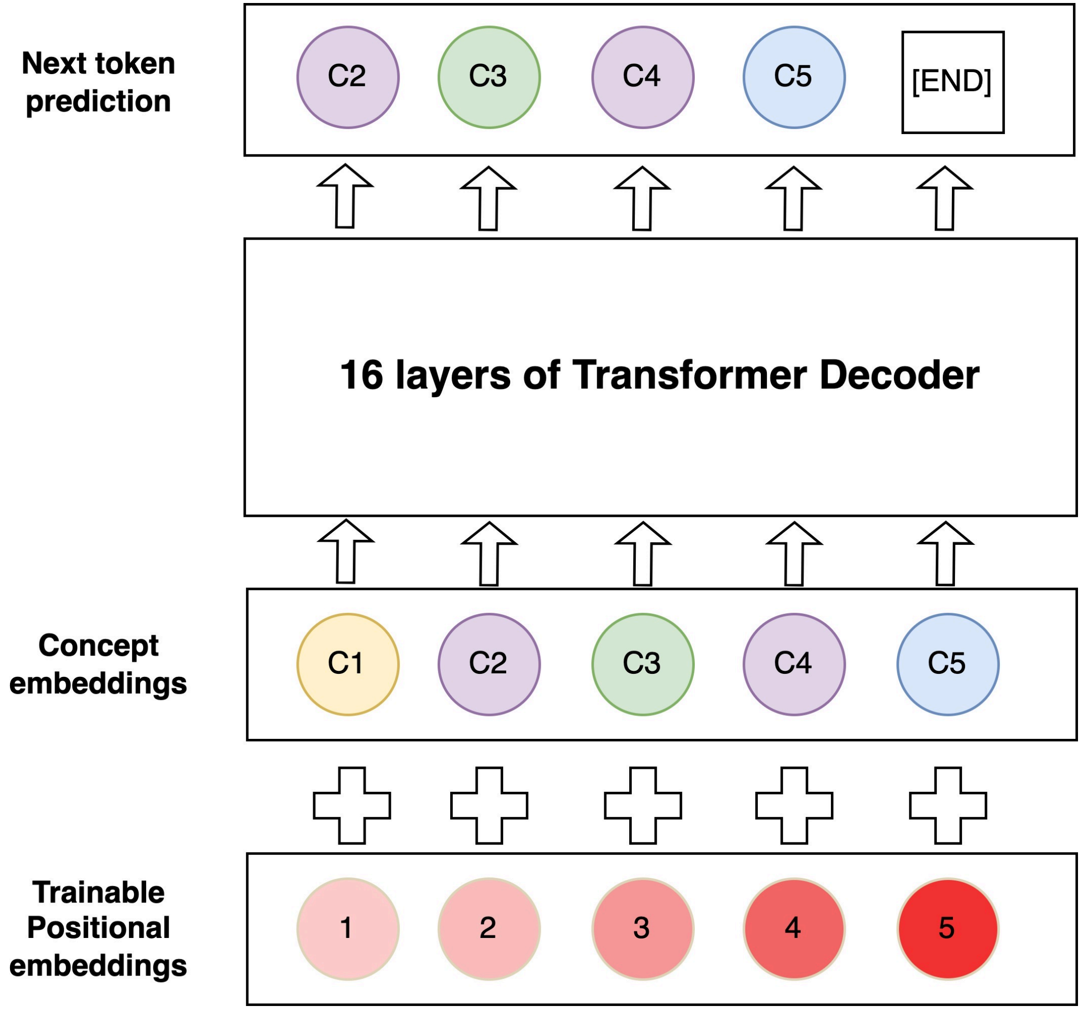


Generative Model





Training a Generative Model



Data Preprocessing

- Condition, drug, procedure
- Context window 512
- Min number of concepts 20
- Truncate the long sequences
- 3 million patients after filtering

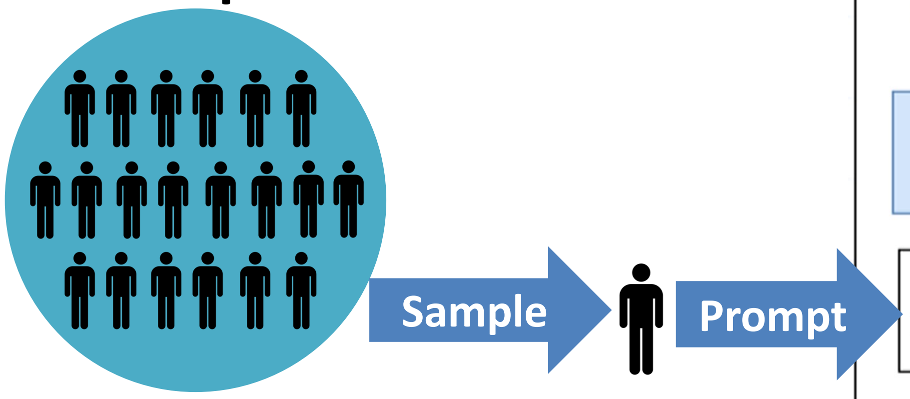
Training parameters

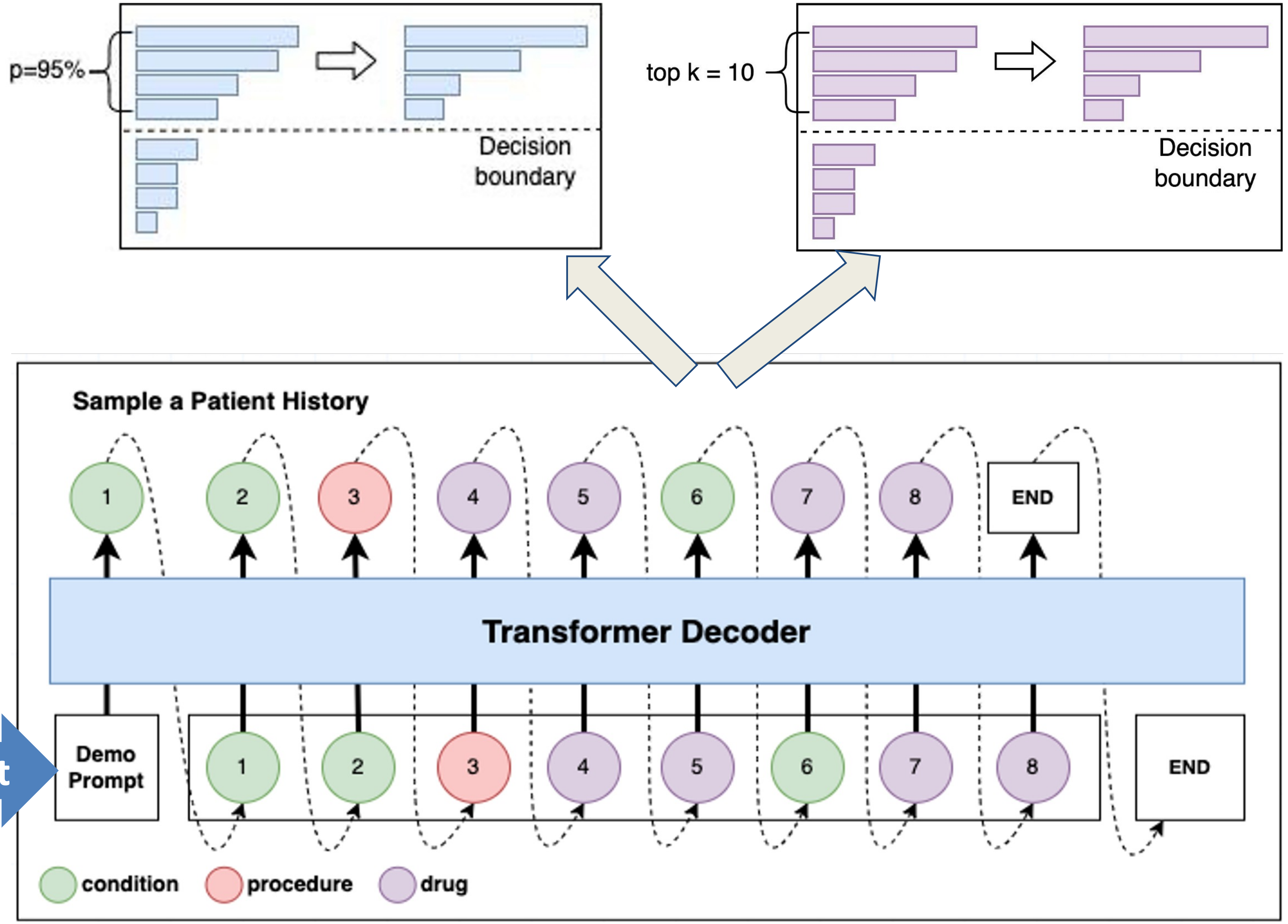
- Batch size 32
- Learning rate 1e-5
- Adam optimizer
- 2 epochs
- Save every 10000 steps

Generate new patient sequences

Inference model
• Top k=100, 200, 300 • Top p=95%, 100% • Generated 500K for each sampling strategy

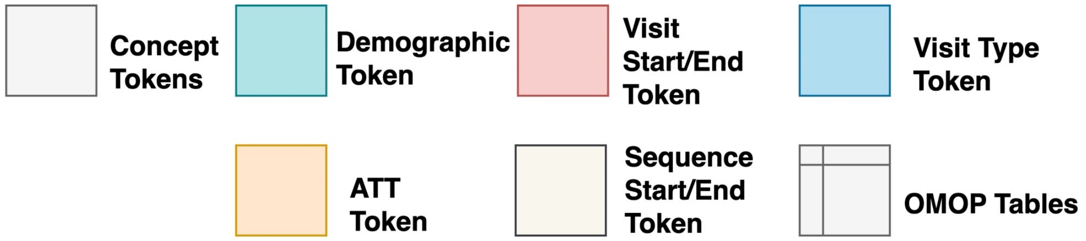
Patient Population



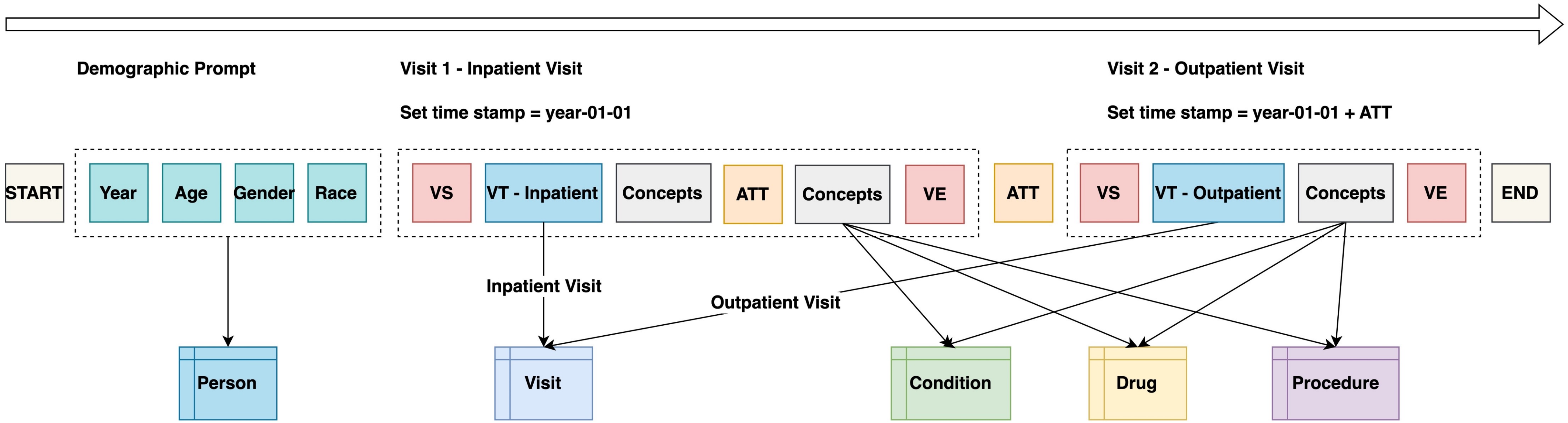


OMOP Converter

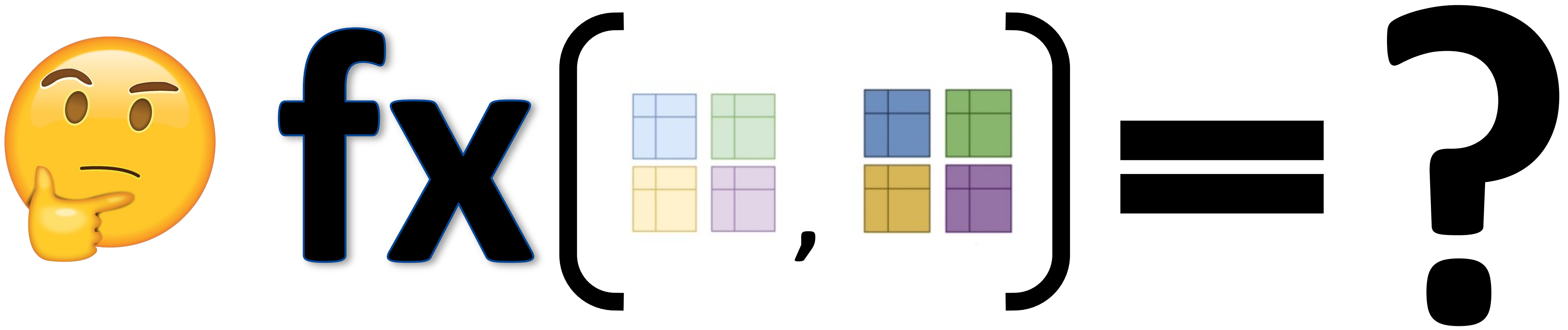
Patient Representation and OMOP Converter



Convert each patient sequence into a set of records in OMOP tables chronologically



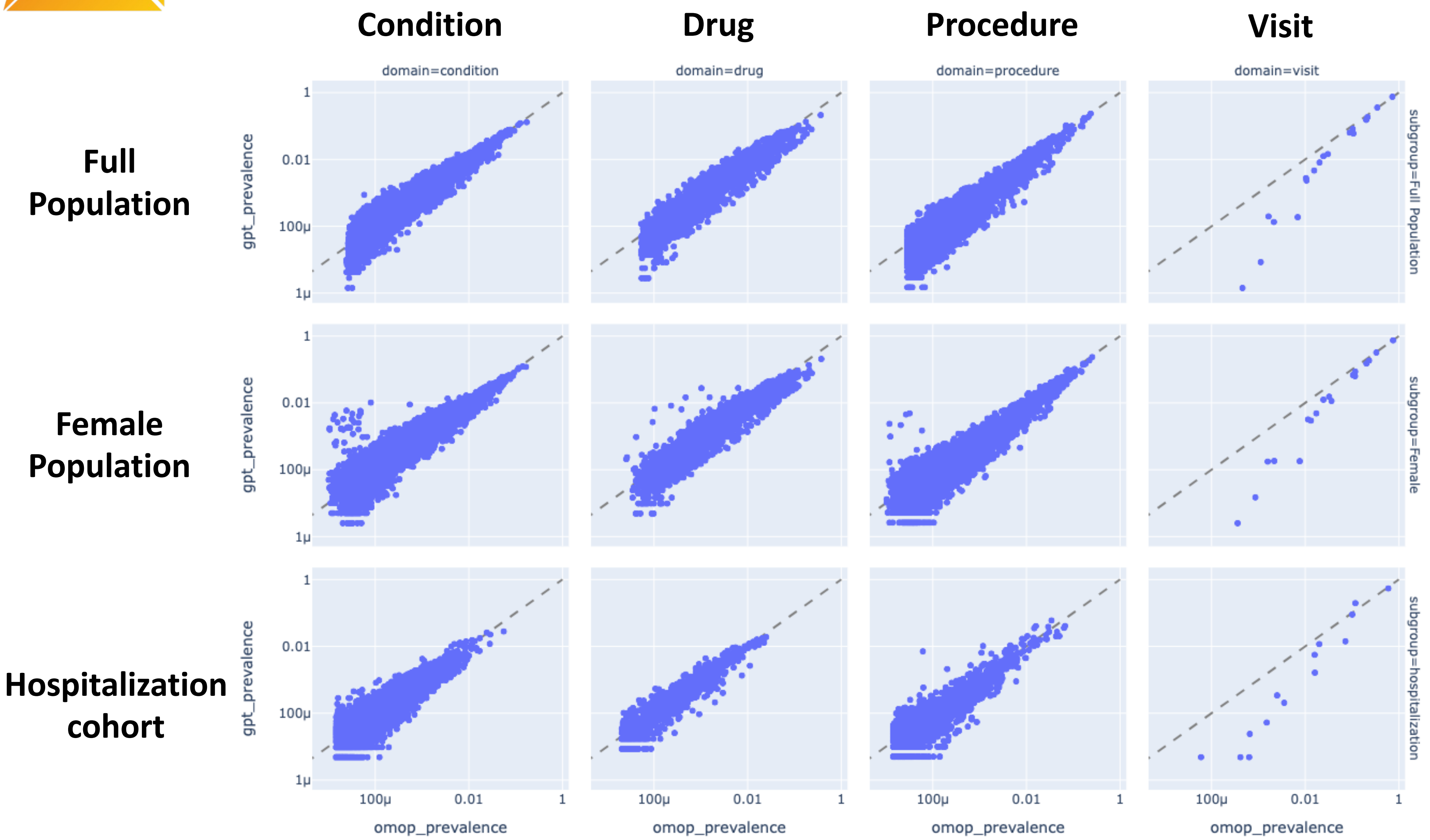
How do you measure the similarity of two OMOP instances?



Evaluation framework

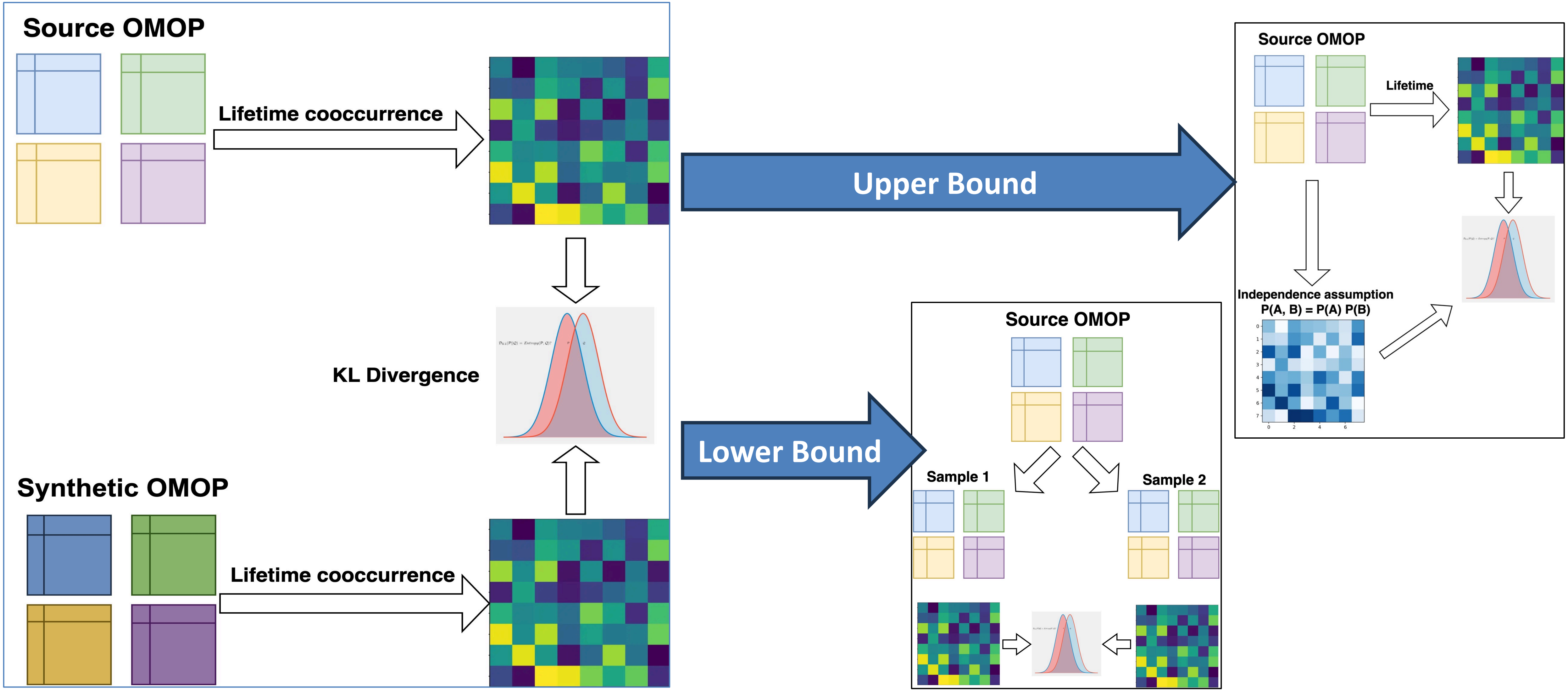
- Level 1: Concept distributions at the full population, subgroups, cohorts. Marginal distribution e.g. P(a; group)
- Level 2: Similarity of co-occurrence matrices at the full population. Conditional distribution e.g. P(a|b)
- Level 3: Logistic regression performance on synthetic cohorts. Proxy for joint distribution e.g. P(a, b, c, d ; group)

Level 1: Concept distributions

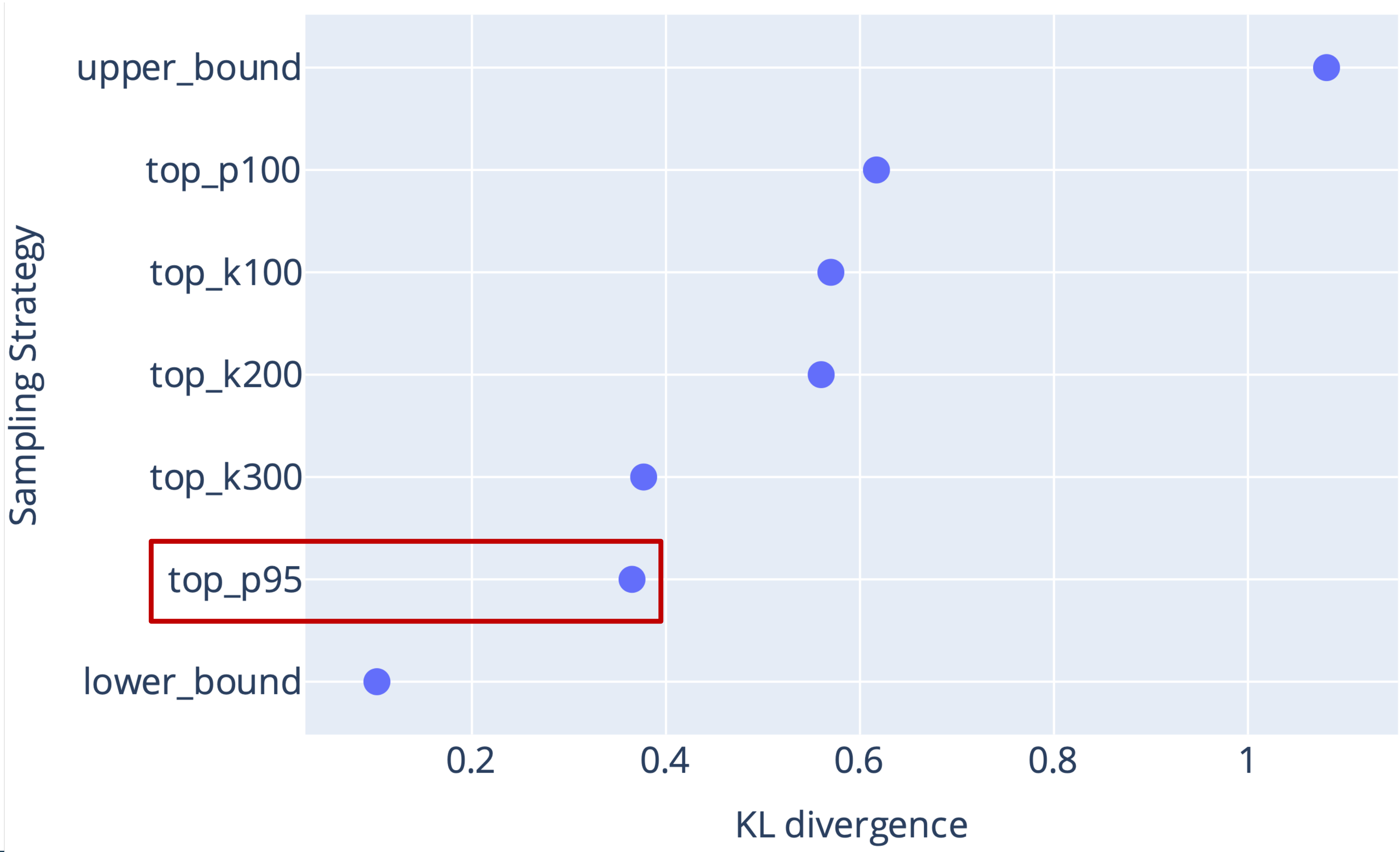


- Synthetic data: Top P=95%
- X: source data
- Y: synthetic data

Level 2: Similarity of co-occurrence matrices



Level 2: Similarity of co-occurrence matrices



- Top k=100, 200, 300
- Top p=95%, 100%
- Sampling strategies affect results.
- Top p=95% has the best KL-divergence

Level 3: Logistic Regression model performance

	Cohort Definition
HF readmission	HF patients who have a 30-day all-cause readmission. Observation window: 360 days, Prediction windows 30 days
Hospitalization	2-year risk of hospitalization starting from the 3rd year since the initial entry into the EHR system Observation window: 540 days, hold-off window: 180 days, Prediction windows 720 days
COPD readmission	COPD patients who have a 30-day all-cause readmission. Observation window: 360 days, Prediction windows 30 days
Afib ischemic stroke	Afib patients with 1 year risk since the initial diagnosis of afib ischemic stroke Observation window: 720 days, Prediction windows 360 day
CAD CABG	Patients initially diagnosed with Coronary Arterial Disease (CAD) without any prior stent graft that will receive the Coronary artery bypass surgery (CABG) treatment Observation window: 720 days, Prediction windows 360 day

Level 3: Logistic Regression model performance

	Real data	Top P=95%	Top P=100%	Top K=100	Top K=200	TOP K=300
HF readmission	Pre = 25.7 AUC = 65.7 PR = 39.3	Pre = 27.6 AUC = 69.2 PR = 45.7	Pre = 28.4 AUC = 65.9 PR = 41.8	Pre = 30.7 AUC = 68.1 PR = 47.8	Pre = 29.3 AUC = 54.0 PR = 32.9	Pre = 26.5 AUC = 64.9 PR = 39.3
Hospitalization	Pre = 5.6 AUC = 75.3 PR = 19.5	Pre = 5.2 AUC = 77.1 PR = 21.4	Pre = 7.3 AUC = 68.3 PR = 16.5	Pre = 2.8 AUC = 87.0 PR = 22.1	Pre = 5.2 AUC = 84.2 PR = 20.8	Pre = 6.3 AUC = 78.7 PR = 24.6
COPD readmission	Pre = 34.5 AUC = 74.2 PR = 83.8	Pre = 37.8 AUC = 76.4 PR = 84.4	Pre = 47.2 AUC = 74.1 PR = 67.2	Pre = 26.4 AUC = 75.9 PR = 90.3	Pre = 28.3 AUC = 70.1 PR = 82.8	Pre = 34.5 AUC = 68.8 PR = 80.2
Afib ischemic stroke	Pre = 8.7 AUC = 84.0 PR = 48.5	Pre = 10.2 AUC = 78.9 PR = 41.2	Pre = 10.4 AUC = 70.7 PR = 39.1	Pre = 16.6 AUC = 77.1 PR = 50.5	Pre = 15.8 AUC =68.9 PR = 36.6	Pre = 10.8 AUC = 76.8 PR = 38.5
CAD CABG	Pre = 7.1 AUC = 88.4 PR = 55.9	Pre = 4.1 AUC = 81.5 PR = 25.2	Pre = 4.4 AUC = 52.9 PR = 4.3	Pre = 7.2 AUC = 75.6 PR = 38.5	Pre = 4.9 AUC = 73.5 PR = 24.3	Pre = 4.0 AUC = 79.0 PR = 24.1

Privacy Evaluation

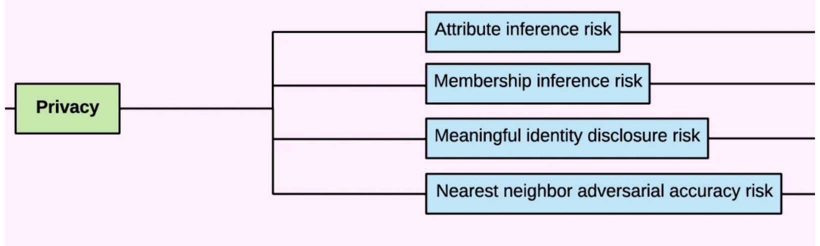
- Membership inference risk: sequence alignment between synthetic and real datasets
- Attribute inference risk: infer the sensitive attributes
- ….

Article | Open access | Published: 09 December 2022

A Multifaceted benchmarking of synthetic electronic health record generation models

Chao Yan, Yao Yan, Zhiyu Wan, Ziqi Zhang, Larsson Omberg, Justin Guinney, Sean D. Mooney & Bradley A. Malin

Nature Communications **13**, Article number: 7609 (2022) | Cite this article



Conclusion

- **First framework** generated longitudinal synthetic EHR data using OMOP CDM.

- Designed an innovative **patient representation**, which allowed the reconstruction of patient medical timeline without loss of temporal information.

- **Comprehensive evaluation procedures** showed that the synthetic data preserved the underlying characteristics of the real patient population.

Acknowledgement

Team

Xinzhuo (Zoey) Jiang
Nishanth Parameshwar Pavinkurve
Krishna S. Kalluri
Elise L. Minto
Jason Patterson
Karthik Natarajan

OHDSI (APOLLO)

Martijn Schuemie
Yong Chen
Egill Fridgeirsson
Chungsoo Kim
Jenna Reps
Marc Suchard
Xiaoyu Wang

DBMI

George Hripcsak
Lingying Zhang
Harry Reyes
Tara Anand
Maura Beaton
Nripendra Acharya

Grants

This project is partially supported by 5U01TR002062 and 5U2COD023196