Community Call Nov 14th
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

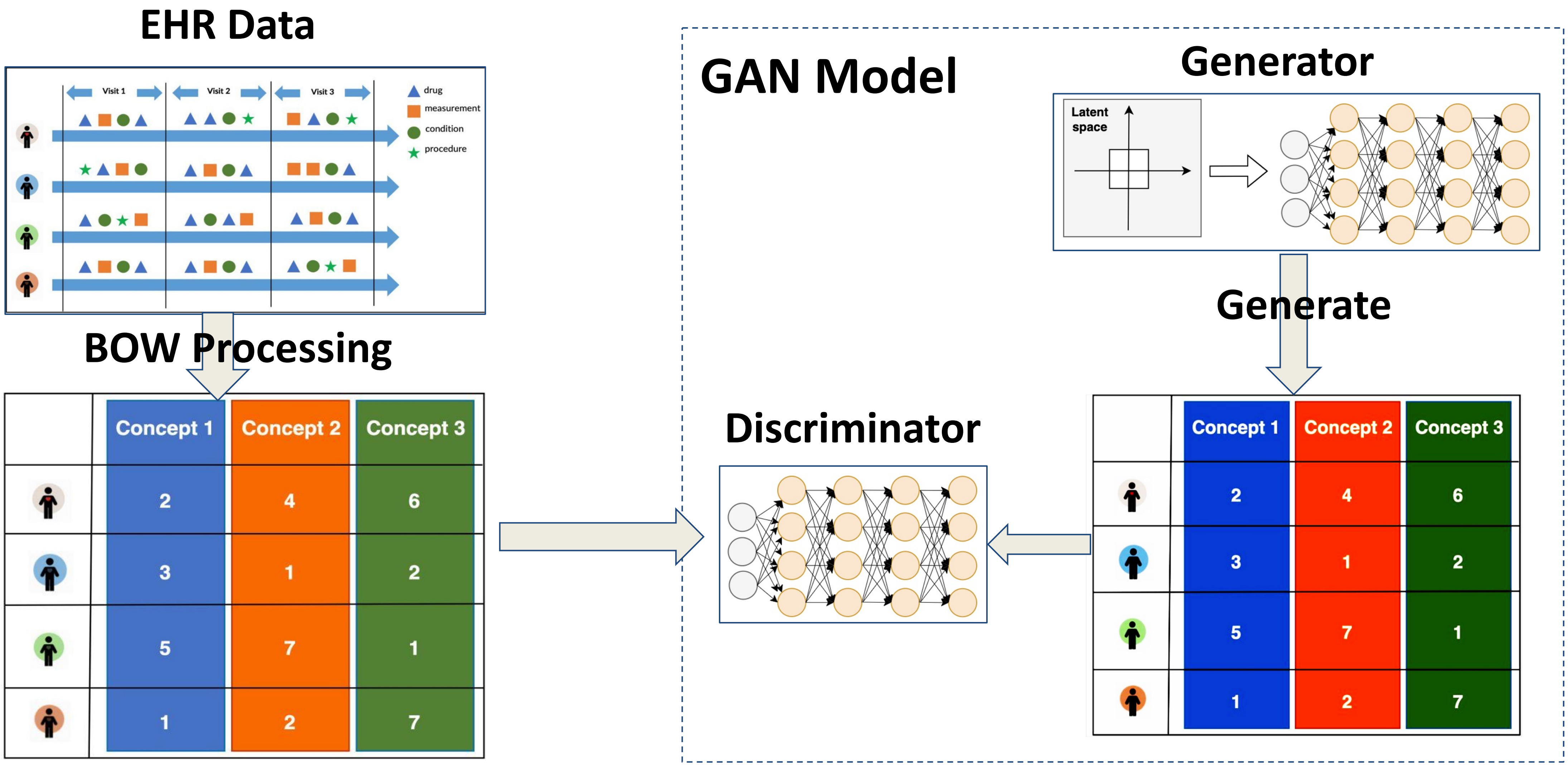
Tool development

Training and education

Fairness and Bias

- Debiasing the source data
- Counterfactual dataset

Common Approach: Bag of Word (BOW) + GAN



JOURNAL ARTICLE

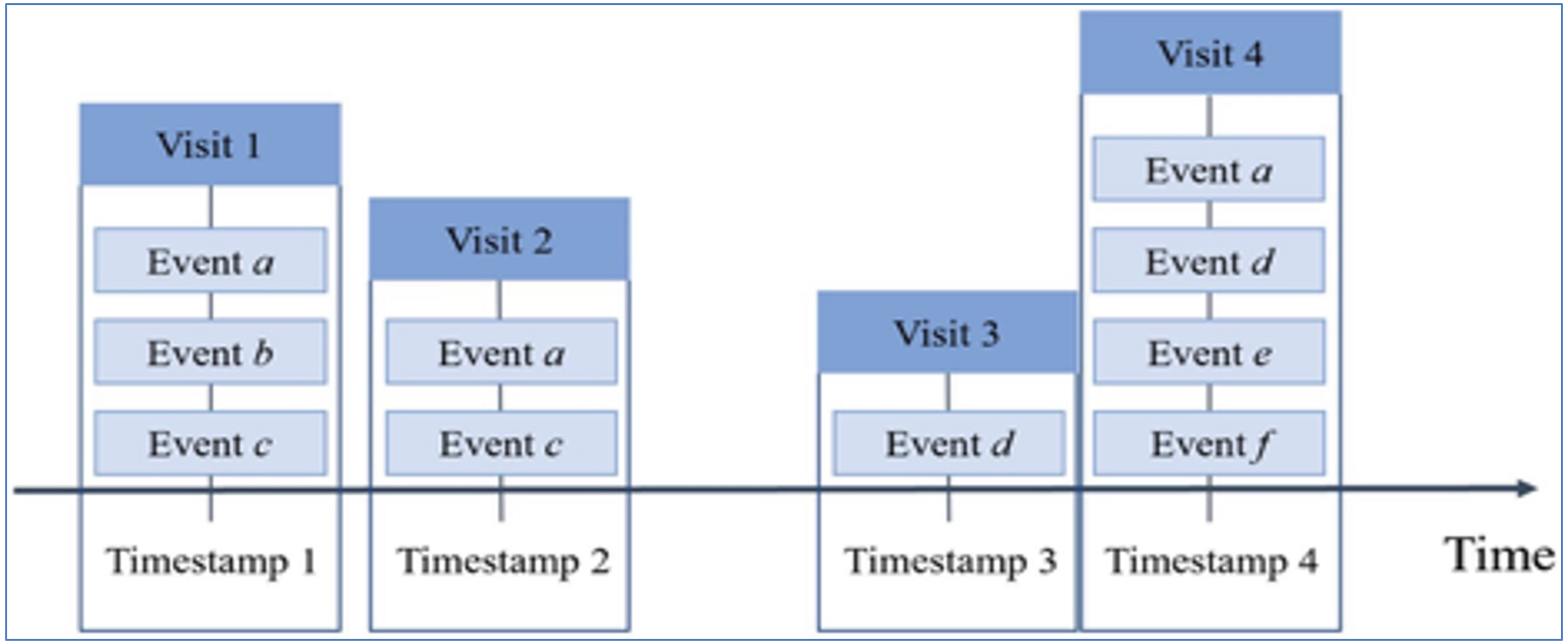
SynTEG: a framework for temporal structured electronic health data simulation FREE

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 ▾

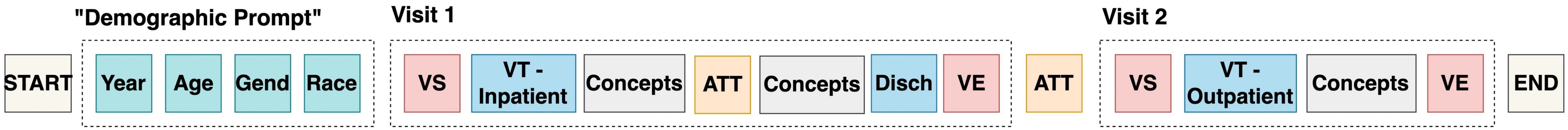


- All visits assume to end on the same day as the visit start (Not true for inpatient visits)

- Visit type is missing
- Discharge type is missing

- Not easily disseminated for use

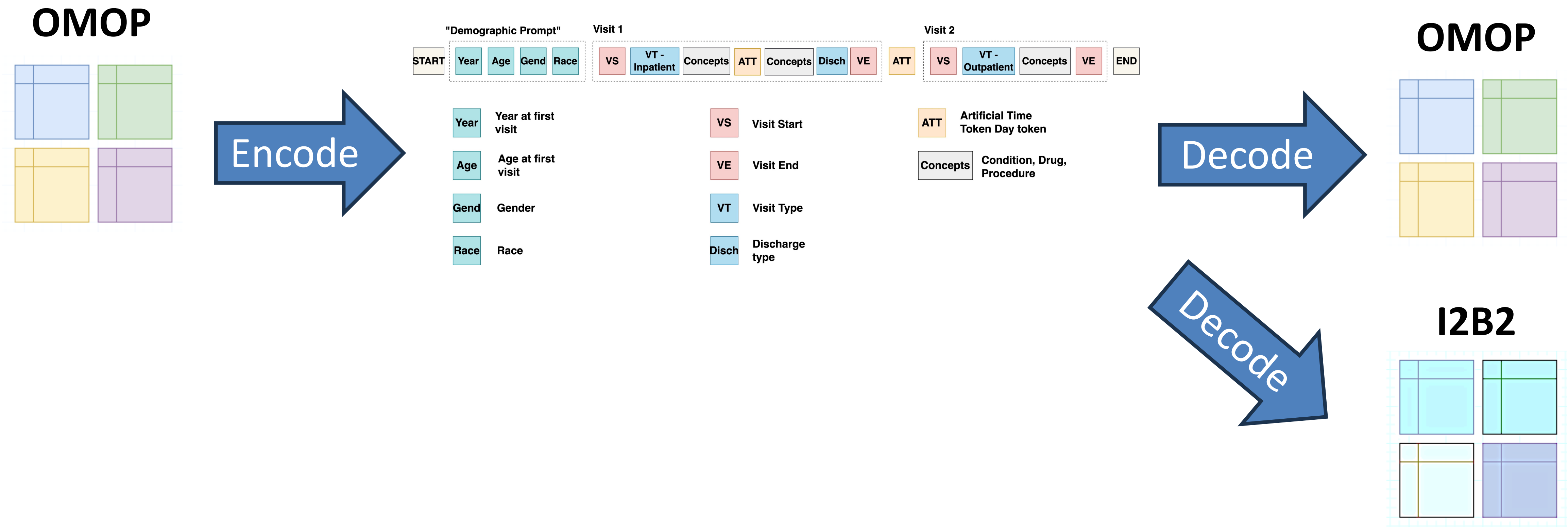
Patient Representation



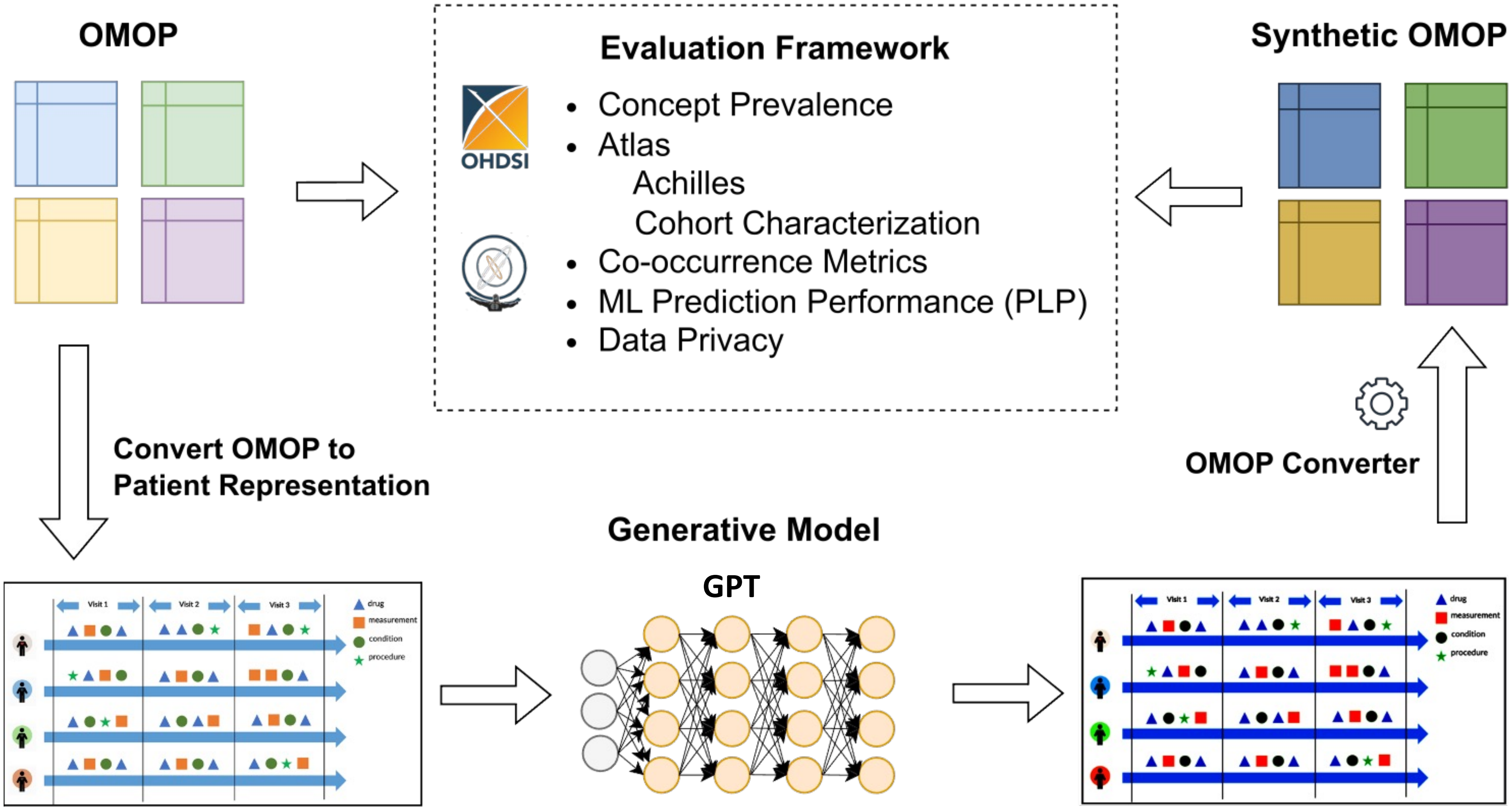
Year	Year at first visit
Age	Age at first visit
Gend	Gender
Race	Race
VS	Visit Start
VE	Visit End
VT	Visit Type
Disch	Discharge type
ATT	Artificial Time Token Day token
Concepts	Condition, Drug, Procedure

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

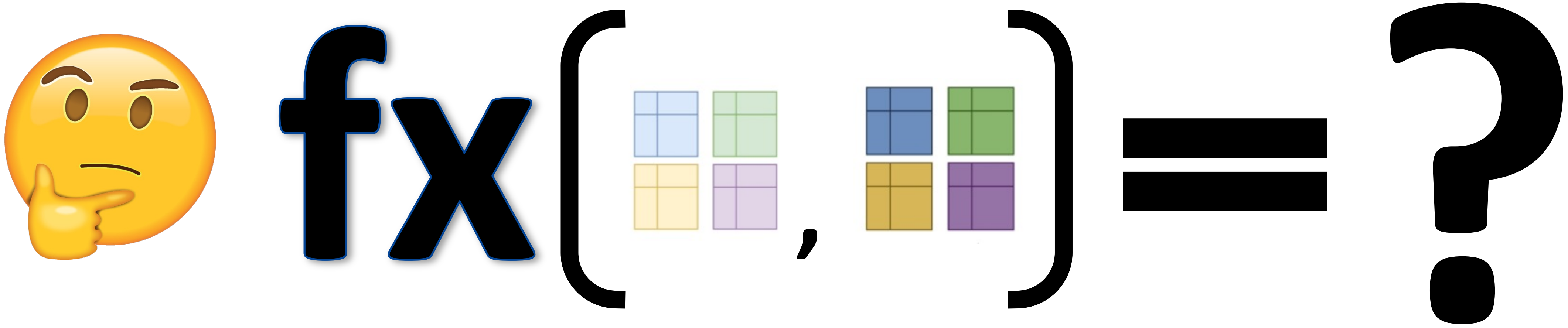
Patient Representation as messenger



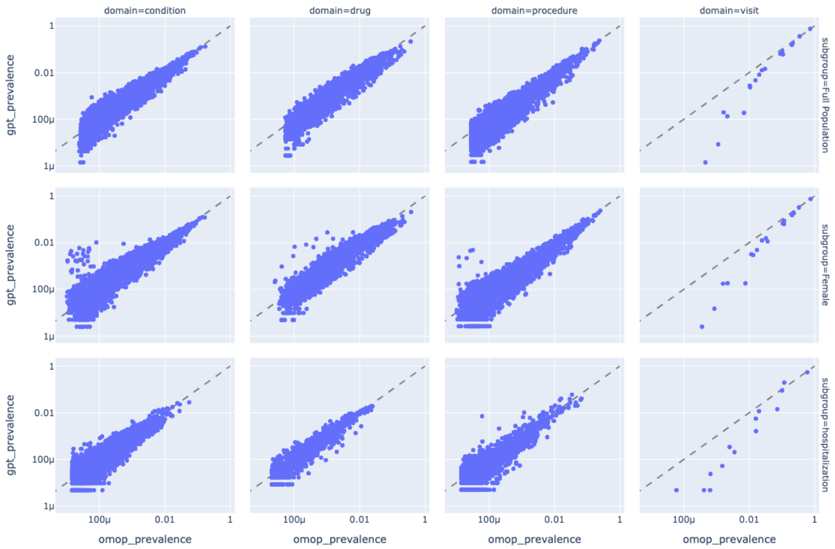
Proposed Synthetic Data Framework



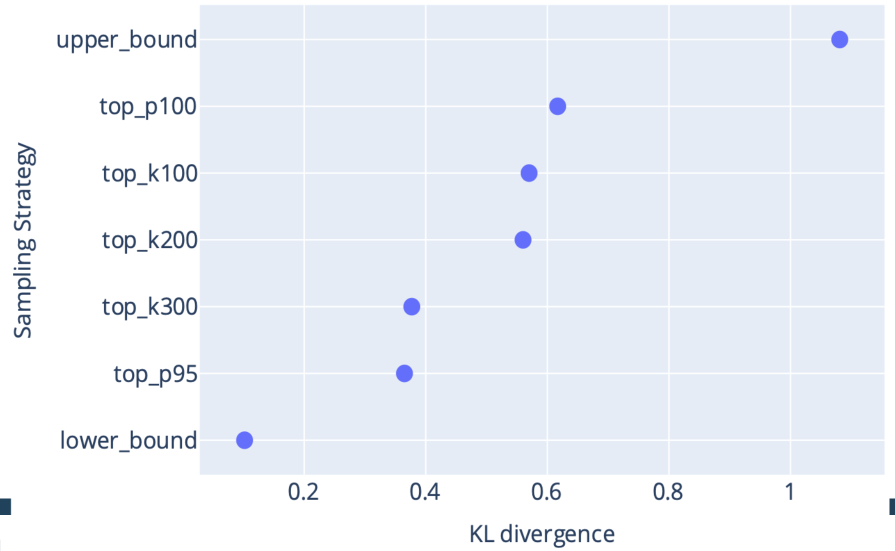
How do you measure the similarity of two OMOP instances?



Level 1: Concept distributions



Level 2: Similarity of co-occurrence



Level 3: Logistic regression performance on synthetic cohorts

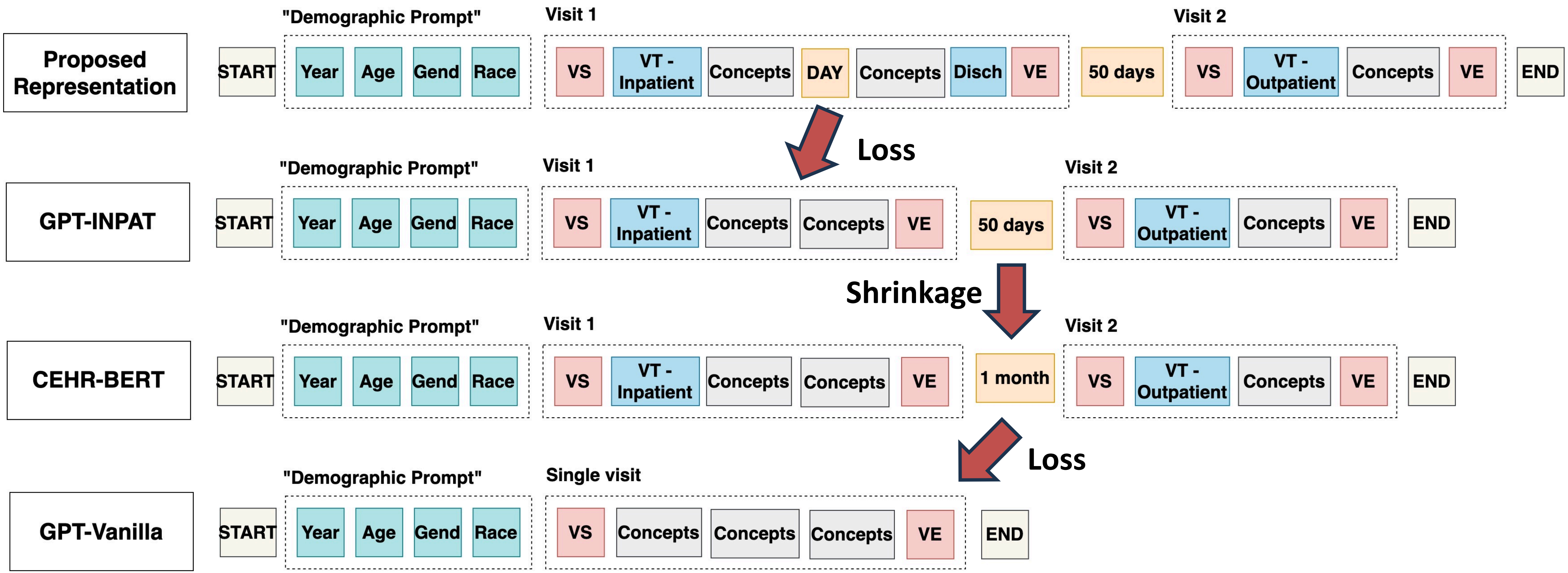
	Real data	Top P=95%	Top P=100%	Top K=100	Top K=200	TOP K=300
HF readmission	Pre = 25.7 AUC = 65.7 PR = 39.3	Pre = 27.6 AUC = 69.2 PR = 45.7	Pre =27.7 AUC = 52.4 PR = 29.0	Pre = 30.7 AUC = 68.1 PR = 47.8	Pre = 29.3 AUC = 54.0 PR = 32.9	Pre = 26.5 AUC = 61.1 PR = 33.8
Hospitalization	Pre = 5.6 AUC = 75.3 PR = 19.5	Pre = 5.2 AUC = 77.1 PR = 21.4	Pre = 7.4 AUC = 71.3 PR = 20.2	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 = 84.7 PR = 31.3	Pre = 4.9 AUC = 73.5 PR = 24.3	Pre = 4.0 AUC = 79.0 PR = 24.1

Loss of Temporal Information (LOTI)

$$LOTI = E_{p(T)}\left[T - G(F(T))\right]$$

- T denotes a time interval
- F denotes the function that generates the ATT token from T
- G denotes the inverse of F that converts ATT back to T’
- G(F(T)) is the reconstructed time interval

Patient Representation Comparison



Loss of Temporal Information (LOTI)

Representation	*Between visit ATT token*	*Between inpatient span ATT token*	*LOTI*
Proposed representation	Day token for $T \leq 1080$ LT token for $T > 1080$	Day token	7.739
GPT-INPAT	Day token for $T \leq 1080$ LT token for $T > 1080$	N/A	7.962
CEHR-BERT	Day token for $T < 7$ Week token for $7 \leq T < 30$ Month token for $30 \leq T < 360$ LT token $T \geq 360$	N/A	31.482
GPT-Vanilla	N/A	N/A	111.164

Time Sensitive Forecasting via MC

$$P(\delta_t|h) \approx \frac{\sum_{i=1}^{n} \mathbb{1}\left[M_{gpt}(h) = \delta_t\right]}{n}$$

Predict the time interval till next visit $E(\delta_t)$

$$P(v|E(\delta_t), h) \approx \frac{\sum_{i=1}^{n} \mathbb{1}\left[M_{gpt}\left(E(\delta_t), h\right) = v\right]}{n}$$

Predict most likely visit type ***V***

$$P(c|v, E(\delta_t), h) \approx \frac{\sum_{i=1}^{n} \mathbb{1}\left[M_{gpt}\left(v, E(\delta_t), h\right) = c\right]}{n}$$

Predict most likely concepts

- ***h*** denotes patient history
- δ_t denotes time interval
- ***v*** denotes visit type
- ***n*** denotes the number of samples

Conclusion

- **First deep learning 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

Columbia DBMI

George Hripcsak

Lingying Zhang

Harry Reyes

Tara Anand

Maura Beaton

Nripendra Acharya

Grants

This project is partially supported by 5U01TR002062 and 5U2COD023196

Thank you!

Email: cp3016@cumc.columbia.edu