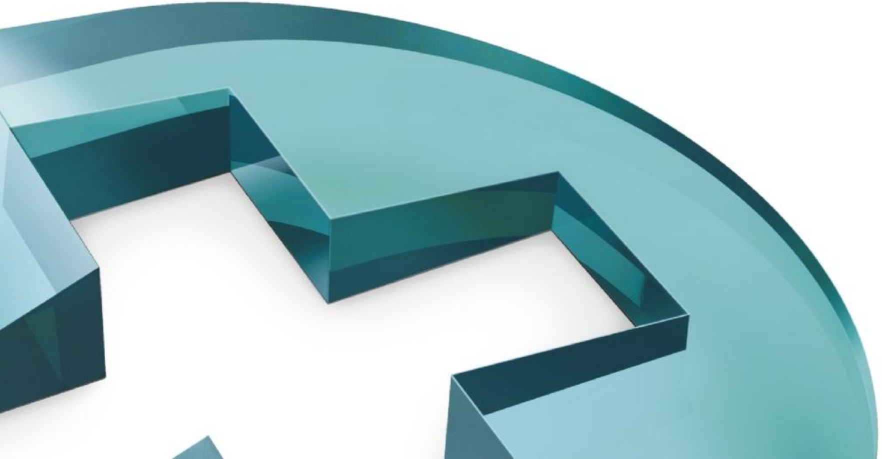
HOSPITAL DA LUZ
LEARNING HEALTH
TRAINING, RESEARCH & INNOVATION CENTER

May 2024 LISBON

Leveraging Synthetic Data for Enhanced Healthcare Analytics

Nuno Machado[1], Jorge Cerejo[2], Simão Gonçalves[2], Bernardo Neves[2], José Maria Moreira[2], Nuno A. da Silva[2]

[1]Instituto Superior Técnico, Universidade de Lisboa, Lisboa, Portugal,
[2]Hospital da Luz Learning Health, Luz Saúde, Lisboa, Portugal



Overview

- Introduction
- Methods
- Results
- Conclusions and Future Work

Introduction

Synthetic Data

- **Artificially generated**
- **Models real data**
- Preserves **privacy** and **confidentiality** of original data
- Useful to **augment** datasets or minority classes in datasets

Introduction

Objectives

- Generate **Synthetic Data** from datasets in the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) format with different models
- Compare the data generated by different models via relevant **metrics**

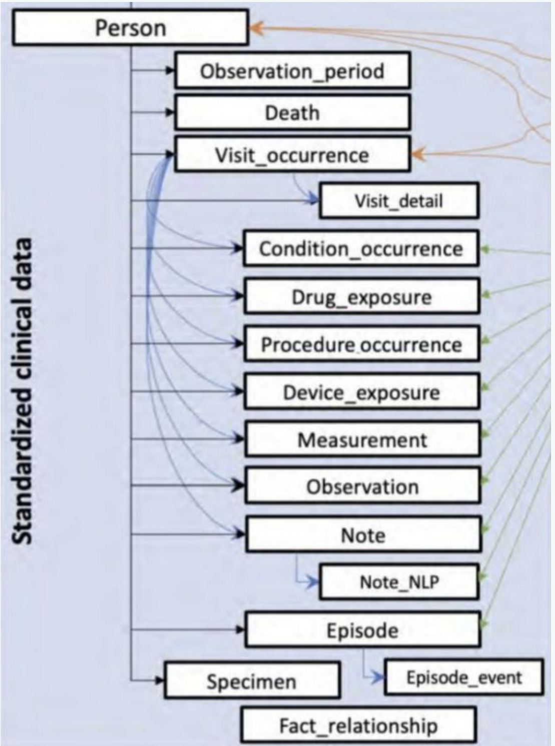


Fig. 1 - OMOP-CDM format

Methods

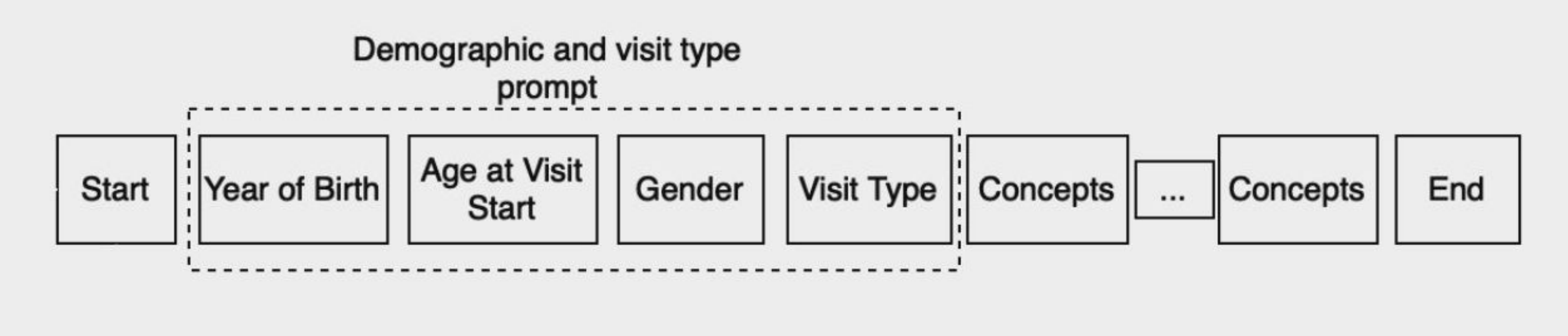


Fig. 2 - Visit Level Representation

Pipeline

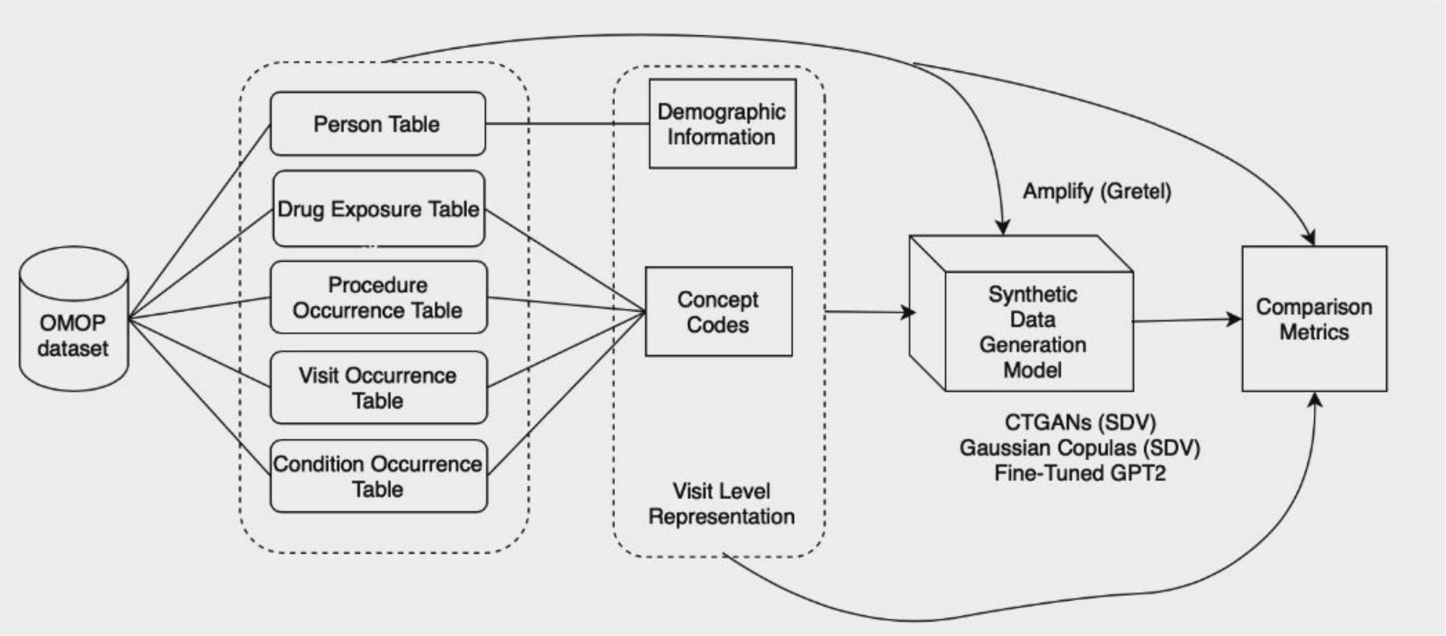


Fig. 3 - Pipeline’s diagram

Methods

Metrics to evaluate synthetic data

- **Prevalence Comparison Plots**
- **Distributions Comparison Plots**
- **Log-Cluster**
- **Kullback-Leibler** (KL) Divergence between co-occurrence matrices of pairs of concepts

Results - Fine Tuned GPT-2 model

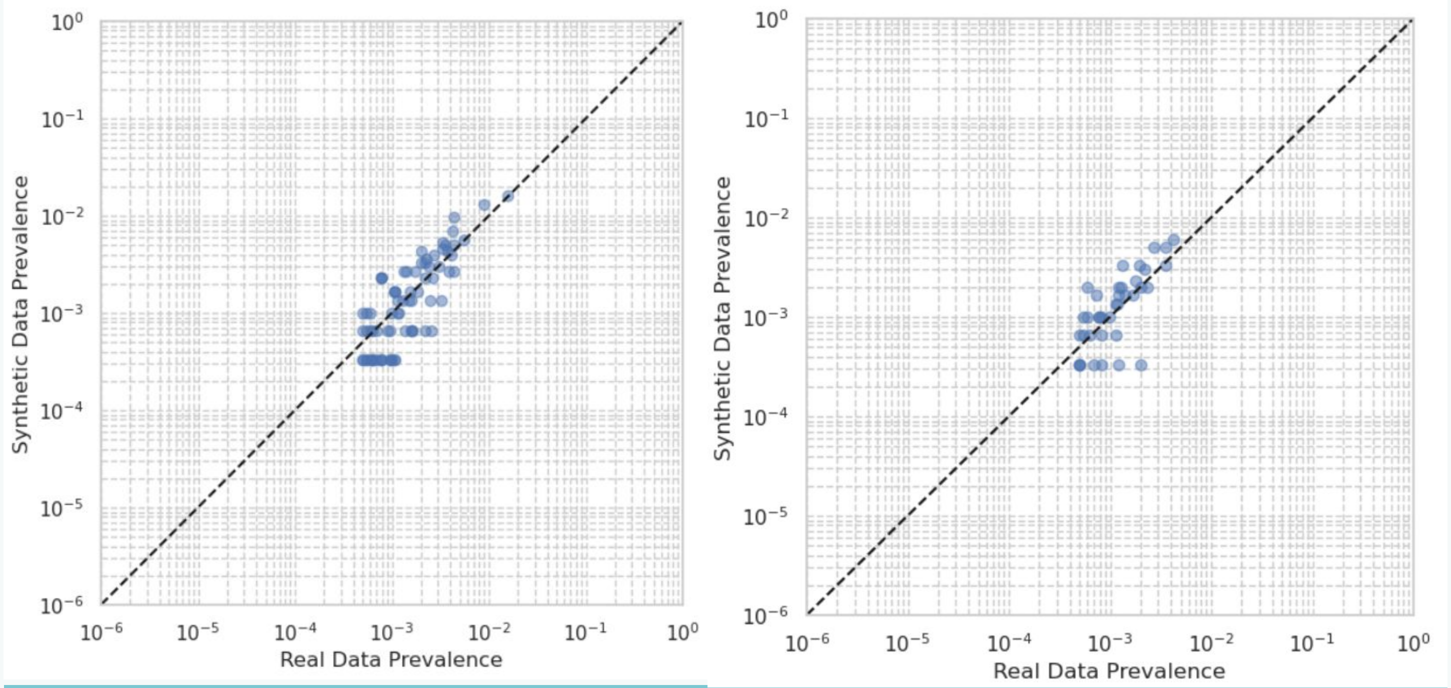


Fig. 4 - Prevalence plot for the drug (left) and condition (right) domains.

Results - Fine Tuned GPT-2 model

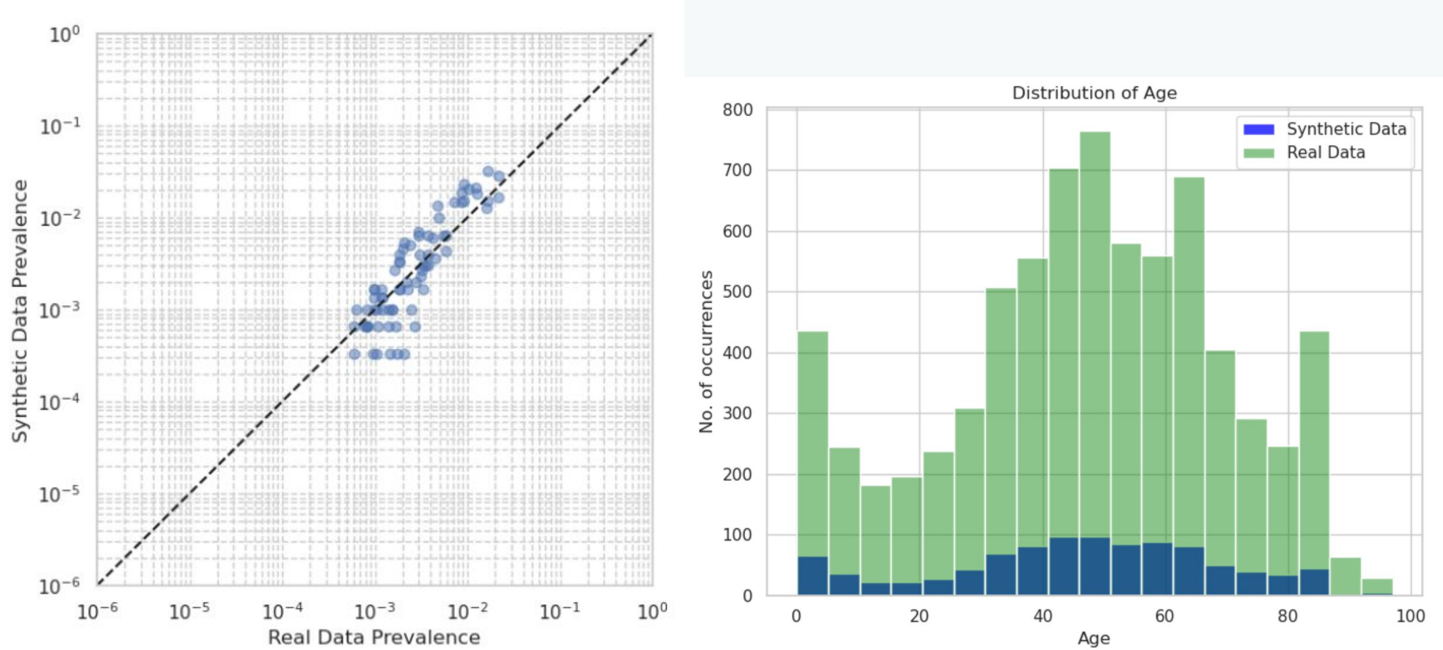
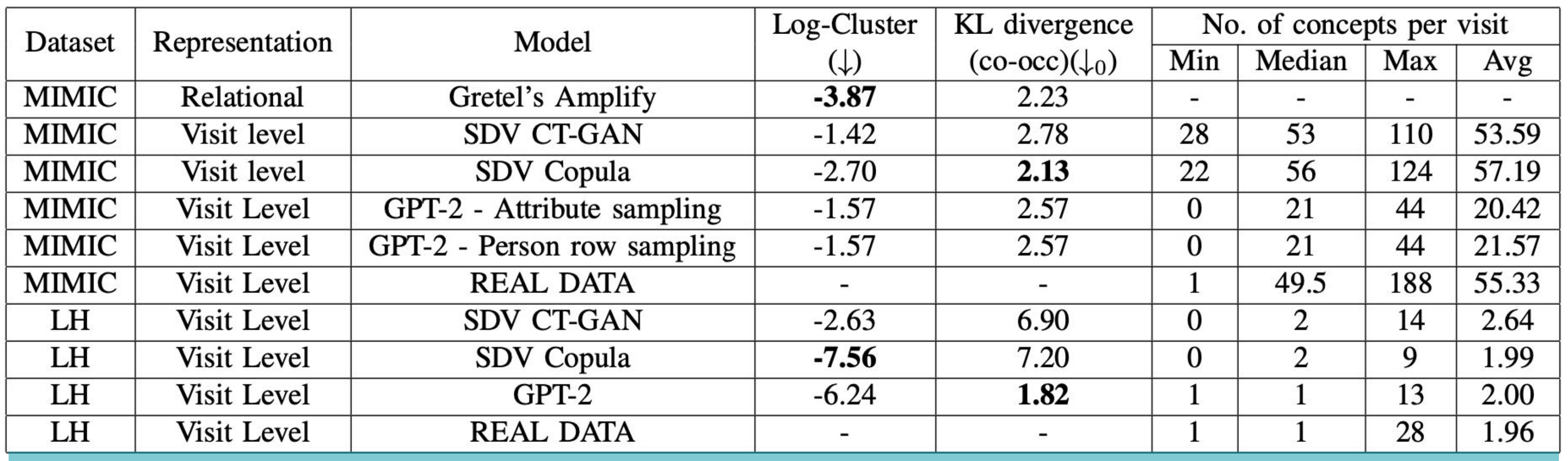


Fig. 5 - Prevalence plot for the procedure domain (left) and age distribution plot (right).

Results



Dataset	Representation	Model	Log-Cluster ($\downarrow$)	KL divergence (co-occ)($\downarrow_0$)	No. of concepts per visit			
					Min	Median	Max	Avg
MIMIC	Relational	Gretel's Amplify	**-3.87**	2.23	-	-	-	-
MIMIC	Visit level	SDV CT-GAN	-1.42	2.78	28	53	110	53.59
MIMIC	Visit level	SDV Copula	-2.70	**2.13**	22	56	124	57.19
MIMIC	Visit Level	GPT-2 - Attribute sampling	-1.57	2.57	0	21	44	20.42
MIMIC	Visit Level	GPT-2 - Person row sampling	-1.57	2.57	0	21	44	21.57
MIMIC	Visit Level	REAL DATA	-	-	1	49.5	188	55.33
LH	Visit Level	SDV CT-GAN	-2.63	6.90	0	2	14	2.64
LH	Visit Level	SDV Copula	**-7.56**	7.20	0	2	9	1.99
LH	Visit Level	GPT-2	-6.24	**1.82**	1	1	13	2.00
LH	Visit Level	REAL DATA	-	-	1	1	28	1.96

Table 1: Metrics comparison between the different models in the different datasets

Conclusions and Future Work

Conclusions

- Generated **synthetic data** with data in the OMOP-CDM format
- Promising results with the **fine-tuned GPT-2 model**
- **Log-Cluster metric** values in line with the literature

Future Work

- Use **concepts descriptions** instead of codes for the fine-tuning
- Experiments with an **LLM Pre-trained on medical concepts**
- Reconvert the visit level representation to the **OMOP-CDM format**