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

Abstract

INTRODUCTION: Synthetic data is gaining recognition in the healthcare sector as a promising approach to address privacy and data scarcity concerns. By augmenting existing datasets or replicating them statistically while preserving patients' privacy, synthetic data offers a solution to these issues [1]. This project aims to explore different methods for generating synthetic data in the OMOP-CDM (Observational Medical Outcomes Partnership Common Data Model) format and assess its quality via relevant metrics.

METHODS: This study utilized a dataset comprising 1000 patients obtained from intelligentCate project in the OMOP-CDM format, which is a standardized open-source database format that uses a standardized medical vocabulary [2]. We developed a visit-level representation of patients' information, where each visit corresponds to a row of data containing all the medical events (see Figure 1A). Our methodology included Generative Adversarial Networks (GANs) and fine-tuned Generative Pre-trained Transformers (GPTs), selected for their ability in handling complex data structures. To assess the quality of the synthetic data generated, we employed two metrics: the Log-Cluster and the Kullback-Leibler (KL) divergence between co-occurrence matrices of pairs of concepts. Additionally, we used prevalence plots for a visual comparison between real and synthetic data prevalence.

RESULTS: The prevalence plot in Figure 1B shows that there is more variability in the frequency of the less common concepts, which is desirable [3]. The results of the Log-Cluster metric (-2.63 and -6.34 for GAN and GPT, respectively) yielded positive outcomes consistent with previous findings reported in the literature, ranging between -1.388 and -10.593 [1]. However, the KL divergence scores obtained differ from those reported in the literature, which range from 0.3 to 0.6 [3]. Specifically, we obtained values of 1.8 (fine-tuned GPT) and 6.9 (GAN).

DISCUSSION & CONCLUSION: Aligning with recent literature, the fine-tuned GPT model demonstrated the most promising results, highlighting transformer-based models as effective tools for generating synthetic data. The results obtained with the fine-tuned GPT model are particularly interesting, considering the utilization of numeric codes, despite the model's main aptitude for textual data. Future work includes exploring the textual information to take advantage of the NLP concepts, evaluate the clinical validity of the generated data and explore pre-trained models in clinical data.

Figure 1

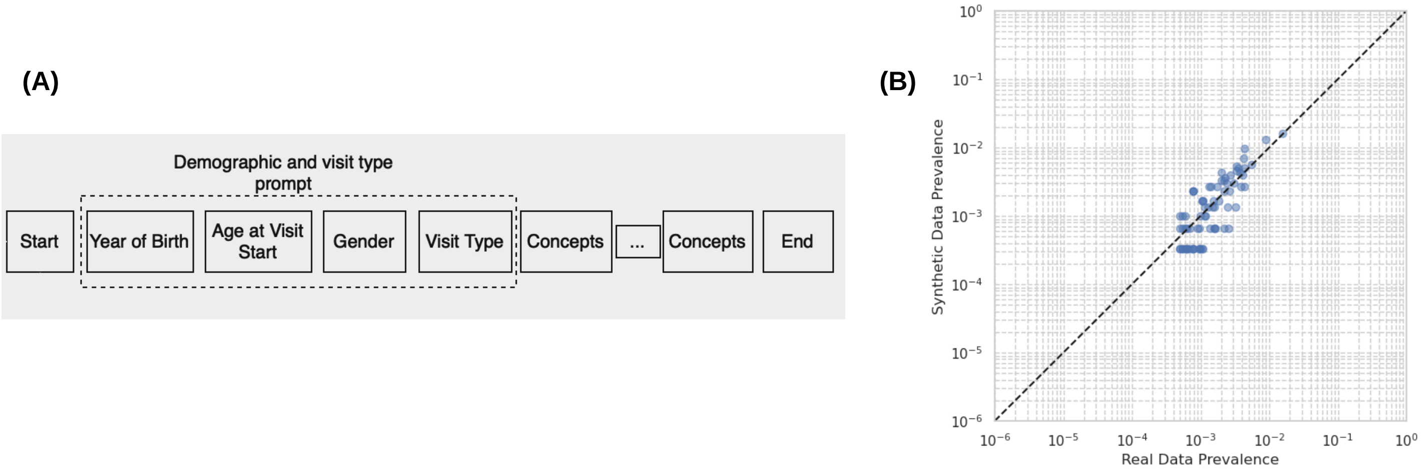

Figure 1: **(A)** Visit-level representation used in the study, where the patient's demographic data and all the medical events within the visit are kept. **(B)** Prevalence plot comparing the synthetic codes generated by the fine-tuned GPT model with the real data in the drug domain. Each data point representing a different concept code.

References

[1] A. Gonzales, G. Guruswamy, and S. R. Smith, "Synthetic data in health care: A narrative review," PLOS Digital Health, vol. 2, p. e0000082, 1 2023.
[2] OHDSI, The Book of OHDSI: Observational Health Data Sciences and Informatics. OHDSI, 2019. [Online]. Available: https://books.google.pt/books?id=JxpnzQEACAAJ
[3] C. Pang, X. Jiang, N. P. Pavinkurve, K. S. Kalluri, E. L. Minto, J. Patterson, and K. Natarajan, "Generating synthetic electronic health records in omop using gpt," 2023. [Online]. Available: https://www.ohdsi.org/2023showcase-503/