An Explorative Study about the Latent Space of Clinical Foundation Models Based on a Common Data Model Database

Min-Gyu Kim[1, 2], Dong Yun Lee[1, 2], Jin Yang Kim[3], Rae Woong Park[1, 2], Joon-Kyung Seong[3, 4]
[1]Department of Biomedical Informatics, Ajou University School of Medicine, Suwon, Republic of Korea
[2]Department of Medical Sciences, Graduate School of Ajou University, Suwon, Republic of Korea
[3]Department of Artificial Intelligence, Korea University, Seoul, South Korea
[4]School of Biomedical Engineering, Korea University, Seoul, South Korea

Background

Recently, there have been researches about clinical foundation models (FMs), which have shown advantages over traditional prediction model. While metrics like F1 score can explain the performance of a model objectively, they are usually inadequate for understanding the internal structure of the model. Also, methods to train such models are still limited to analogies from the language domain.
There are many methods available that enable model understanding, such as visualizing self-attention of each layer or dimension reduction in the latent space. In this study, we aim to understand how we should train clinical foundation models by first training a model using our own data based on OMOP-CDM and visualizing the latent space of the trained model.

Methods

We trained a transformer model based on the bidirectional transformer (BERT) architecture, using data from Ajou university hospital standardized to the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM). Patient records were first translated into a time series format. Additional information such as patient age and gender were prepended to the input series as separate tokens. To provide a better understanding about the domains defined by OMOP-CDM, each token was added to the embedding about its domain, i.e. condition, drug, measurement.
The model was trained using masked language modeling. 15% of the tokens were randomly masked and the model predicted the original tokens. 1% of the total training data was randomly selected, and the CLS tokens of the sample were calculated. The tokens were then reduced to seven dimensions using Uniform Manifold Approximation (UMAP) and clustered with Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN). The result was visualized using t-distributed stochastic neighbor embedding (t-SNE) by reducing to a 2-dimensional plane. The resulting visualization was inspected, and cluster formation was manually evaluated using Term Frequency-Inverse Document Frequency (TF-IDF).

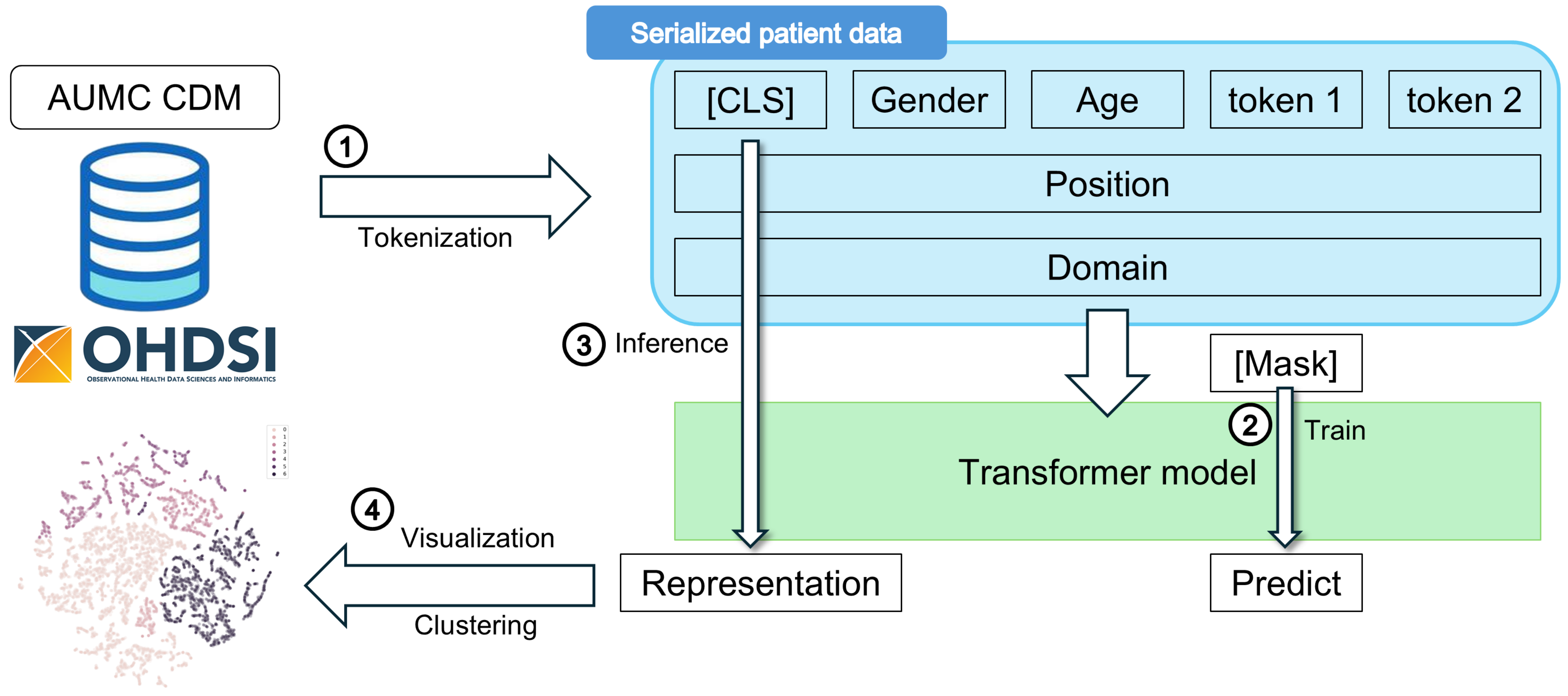

Figure 1. Study flow. First, the OMOP-CDM in Ajou university medical center was transformed into serial data according to each patient, including basic patient information such as gender and age. The data was then fed through a transformer model,

Results

Training loss converged and the model with the least validation error was selected. The clusters were not immediately recognizable with the IDs only, but some was specific enough to make weak assumptions about the cluster. For example, cluster 5 had measurements related to health screening.
The visualization of clusters using representative tokens showed better results in cluster membership. While some tokens representing a cluster was not present for most of the patient data within that cluster, certain tokens clearly showed patterns of grouping (Figure 2), closely resembling the distribution of the cluster.

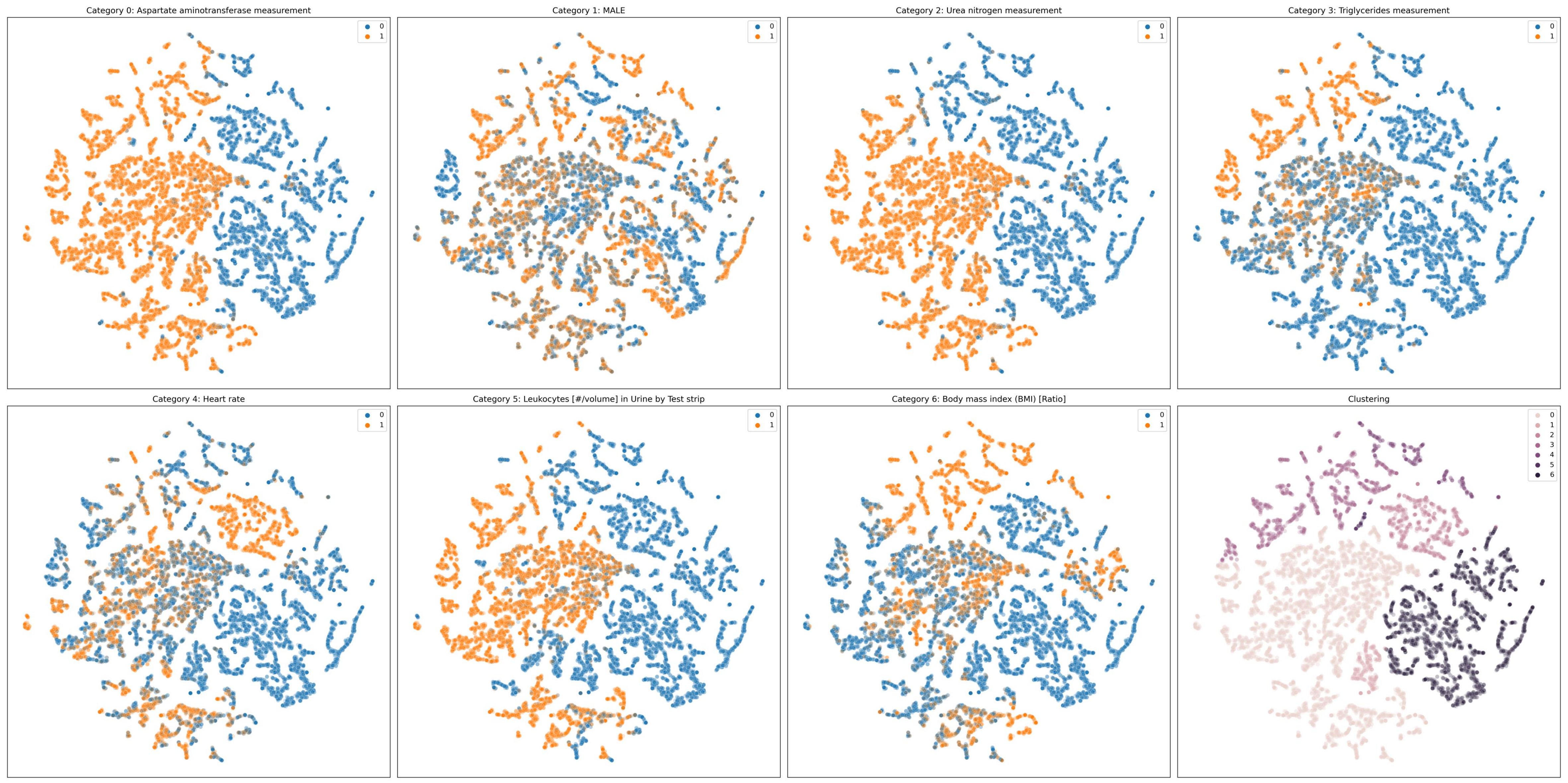

Figure 2. Top 1 representative token of each cluster visualized. (Blue) Patients without representative concept ID of cluster 0 to 6. (Orange) Patients with representative concept ID of cluster 0 to 6.

Conclusions

In this study, we trained a BERT-based clinical foundation model using data from electronic health record converted to OMOP-CDM. The latent space was visualized using dimension reduction techniques and clusters with explainable characteristics were found in some cases. A better optimized approach with different architectures or training method may lead to a better intuitive understanding about the data contained using OMOP-CDM.

Acknowledgement

This research was funded a grant from the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number: HR16C0001) and this research was supported by a Government-wide R&D Fund project for infectious disease research (GFID), Republic of Korea (grant number: HG22C0024, KH124685).

Contact: manjmin6@gmail.com / Min-Gyu Kim