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 are models that can embed patient health records as a single patient representation. These models have shown advantages over traditional prediction models such as improved predictive accuracy, less need for labeled data and easier deployment. For foundation models to be widely accepted, they have to be applicable across different health records systems. However, studies related to clinical FMs are mostly not immediately transferable. Furthermore, 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.[1, 2]

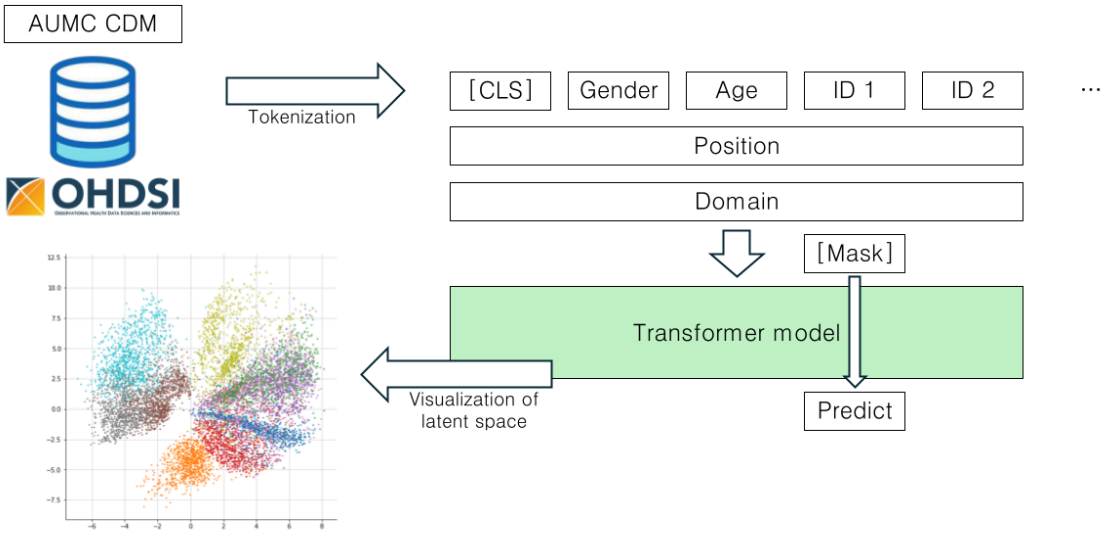
Methods

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). Concept IDs with less than 2,000 appearances total were dropped and the remainder were added to the vocabulary. Using this vocabulary, 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. Maximum length of 512 tokens were applied by randomly selecting the starting index and slicing after 512 tokens. To provide a better understanding about the domains defined by OMOP-CDM, each token was added 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. Cross-entropy was used as the loss function and training was stopped after the loss converged. Embeddings for each token in the vocabulary was calculated and reduced to two dimensions using t-distributed stochastic neighbor embedding (t-SNE) and primary component analysis (PCA). The resulting visualization was inspected, and cluster formation was manually evaluated.

Results

Training loss converged at 1.4, and the model with the least validation error was selected. While in general the embeddings were normally distributed, some clusters were formed. While most of the clusters were not related to each other, some closely related clusters were found. Condition tokens were



more consistently aggregated compared to drug tokens, but drug tokens appeared in similar positions to the related condition.

The overall distribution across all domains was close to a normal distribution, which may suggest underfitting. Special tokens such as gender were also normally distributed. Specifically, year-of-birth did not show a series-like pattern which is usually observed in special tokens of the language domain.

Conclusion

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. While some clustering was observed, most of the distribution did not show an explainable pattern, even when the loss converged. This may suggest the need for a more optimized approach to enable representation of patient information based on 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).

References

1. Wornow, M., Xu, Y., Thapa, R., Patel, B., Steinberg, E., Fleming, S., ... & Shah, N. H. (2023). The shaky foundations of large language models and foundation models for electronic health records. *npj Digital Medicine*, *6*(1), 135.
2. Pang, C., Jiang, X., Kalluri, K. S., Spotnitz, M., Chen, R., Perotte, A., & Natarajan, K. (2021, November). CEHR-BERT: Incorporating temporal information from structured EHR data to improve prediction tasks. In *Machine Learning for Health* (pp. 239-260). PMLR.