Med-BERT v2: clinical foundation model on standardized secondary clinical data

Laila Rasmy[1], Yan Chu[1], Bingyu Mao[1], Khush Patel[1], Zhao Li[1], Hao Yan[1], Ziqian Xie[1], Wenjin Zheng[1], Hua Xu[1], and Degui Zhi[1]
[1] University of Texas Health Science Center in Houston

Background.

Deep learning (DL) based predictive models from electronic health records (EHR) deliver impressive performance in many clinical tasks. The need for large training cohorts, however, are often required, hindering the adoption of DL-based models in scenarios with limited training data size. In our previous work[1], we showed that Med-BERT (https://github.com/ZhiGroup/Med-BERT) trained on patients diagnoses data in standard ICD codes from more than 20 million patients' EHR substantially improves the prediction accuracy for tasks with small cohorts. Med-BERT improved the discriminative accuracy of tasks with fine-tuning training sets of a few hundred samples boosting the AUC by more than 20% or equivalent to the AUC of 10 times larger training sets. Adding more patient information including medications and procedures are known to further increase the prediction accuracy for many clinical tasks[2,3]. However, when we use the earlier version of Med-BERT trained on diagnoses information alone and add to it randomly initialized embeddings for medications and procedures, the model performance deteriorates. Therefore, we trained a new version of Med-BERT adding medications and procedures data. Additionally, we compared the performance of the Med-BERT model trained on claims data (MBv2-Claims) versus the model originally trained on EHR data (MBv2-EHR) to evaluate the generalizability of our approach as well as the generalizability of the pre-trained model.

Methods.

We expanded the input features for the Med-BERT previously trained cohort to cover more than 130,000 features representing different diagnoses in ICD-9 and ICD-10 codes, procedures in CPT, HCPCS, and ICD PCS codes, and medications in Multum ID and Multum categories. Additionally we extracted a similar cohort of around 51 million patients from the US-based Optum de-identified Optum Clinformatics® Data Mart (Optum CDM) claims dataset. For comparison purposes, we randomly selected 20 Million patients from Optum CDM, to train MBv2-Claims versions. We compared the performance of MBv2-EHR and MBv2-Claims against our very first version of Med-BERT(MBv1) on the same downstream disease prediction tasks: the prediction of heart failure among patients with diabetes (DHF) and the prediction of onset of pancreatic cancer (PaCa). We use bidirectional gated recurrent neural networks (Bi-GRU) as our baseline model, as it was the one associated with the best performance in our earlier study. We evaluated the performance boost using different versions of Med-BERT using the area under the receiver operating characteristic curve (AUROC) as our main evaluation metric.

Results.

The discriminative accuracy of the models measured by the AUROC (Table 1), showed a slight increase on the DHF task after adding procedures and medications data (0.2%) while the impact was higher on the PaCa task (3%). The use of our early version of Med-BERT (MBv1) highly boosted the predictive model accuracy, even with using diagnoses data alone, by 2.4% to 3.1% , compared to the baseline Bi-GRU model trained on all data categories. However, when we fine-tuned Bi-GRU on top of MBv1 after adding procedures and medication codes, the performance boost was less than 1%. Newly trained MBv2 using US-based Cerner Health Facts® de-identified EHR-derived dataset (MBv2-EHR) bring the performance boost back up by 3% to 5.2%, and an MBv2 model trained on equal sized population from Optum CDM data boosted the performance by 2.1 and 4.4%.

Table 1. Downstream tasks performance in AUROC(%) for different models

Model	**DHF** *n=50,750*	**PaCa** *n=19,250*
Bi-GRU (diagnosis data only)	82.8	76.1
Bi-GRU (diagnosis + procedures+ medications)	83.0	79.1
Bi-GRU (diagnosis data only) + Med-BERT(pretrained on diagnoses data only)	85.4	82.2
Bi-GRU (diagnosis + procedures + medications) + Med-BERT(pretrained on diagnoses data only)	83.6	79.9
Bi-GRU (diagnosis + procedures + medications) + MBv2-EHR	86.0	84.3
Bi-GRU (diagnosis + procedures + medications) + MBv2-Claims	85.1	83.5

Conclusion.

Adding new features during the finetuning phase will decrease the magnitude of the performance boost offered by the pretrained model, therefore there is a need to pretrain foundation models such as Med-BERT on more clinical data categories. MBv2-Claims provided a comparable performance boost to the MBv2-EHR, even when the downstream tasks only used EHR data.

References

1 Rasmy L, Xiang Y, Xie Z, *et al.* Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ Digit Med* 2021;**4**. doi:10.1038/S41746-021-00455-Y

2 Rasmy L, Wu Y, Wang N, *et al.* A study of generalizability of recurrent neural network-based predictive models for heart failure onset risk using a large and heterogeneous EHR data set. *J Biomed Inform* 2018;**84**:11–6.

3 Rasmy L, Nigo M, Kannadath BS, *et al.* CovRNN—A recurrent neural network model for predicting outcomes of COVID-19 patients: model development and validation using EHR data. *medRxiv* 2021;:2021.09.27.21264121.