Pretrained Patient Trajectories for Adverse Drug Event Prediction Using Common Data Model-based Electronic Health Records

Junmo Kim[1], Joo Seong Kim[2], Ji-Hyang Lee[3], Min-Gyu Kim[4,5], Taehyun Kim[6], Chaeeun Cho[7], Rae Woong Park[4,5,*], Kwangsoo Kim[8,9,*]

[1]Interdisciplinary Program in Bioengineering, Seoul National University, Seoul, Republic of Korea

[2]Division of Gastroenterology, Department of Internal Medicine, Dongguk University Ilsan Hospital, Dongguk University College of Medicine, Goyang, Republic of Korea

[3]Drug Safety Center, Seoul National University Hospital, Seoul, Republic of Korea

[4]Department of Biomedical Sciences, Ajou University Graduate School of Medicine, Suwon, Republic of Korea.

[5]Department of Biomedical Informatics, Ajou University School of Medicine, Suwon, Republic of Korea.

[6]College of Medicine, Hanyang University, Seoul, Republic of Korea

7Department of Medicine, Korea University College of Medicine, Seoul, Republic of Korea

[8]Department of Transdisciplinary Medicine, Institute of Convergence Medicine with Innovative Technology, Seoul National University Hospital, Seoul, Republic of Korea

[9]Department of Medicine, College of Medicine, Seoul National University, Seoul, Republic of Korea

***Corresponding Authors**

Rae Woong Park, PhD

Department of Biomedical Sciences, Ajou University Graduate School of Medicine, Suwon, Republic of Korea.

Department of Biomedical Informatics, Ajou University School of Medicine, Suwon, Republic of Korea.

Address: 164 World cup-ro, Yeongtong-gu, Suwon, Republic of Korea

Email: veritas@ajou.ac.kr

Kwangsoo Kim, PhD

Department of Transdisciplinary Medicine, Institute of Convergence Medicine with Innovative Technology, Seoul National University Hospital

Department of Medicine, College of Medicine, Seoul National University

Address: 101 Daehak-ro, Jongno-gu, Seoul, Republic of Korea

Email: kksoo716@gmail.com

Abstract

Background. Pretraining electronic health record (EHR) data using language models by treating patient trajectories as natural language sentences has enhanced performance across various medical tasks. However, EHR pretraining models have never been utilized in adverse drug event (ADE) prediction. We constructed and externally validated the EHR pretraining model for several ADE prediction tasks and qualitatively analyzed the important features of each ADE cohort.

Methods. A retrospective study was conducted on observational medical outcomes partnership (OMOP)-common data model (CDM) based EHR data from two separate tertiary hospitals. The data included patient information in various domains such as diagnosis, prescription, measurement, and procedure. For pretraining, codes were randomly masked, and the model was trained to infer the masked tokens utilizing preceding and following history. In this process, we adopted domain embedding (DE) to provide information about the domain of the masked token, preventing the model from finding codes from irrelevant domains. For qualitative analysis, we identified important features using the attention matrix from each finetuned model.

Results. 510,879 and 419,505 adult inpatients from two separate tertiary hospitals were included in internal and external datasets. EHR pretraining model with DE outperformed all the other baselines in all cohorts. For feature importance analysis, we demonstrated that the results were consistent with priorly reported background clinical knowledge. In addition to cohort-level interpretation, patient-level interpretation was also available.

Conclusions. EHR pretraining model with DE is a proper model for various ADE prediction tasks. The results of the qualitative analysis were consistent with background clinical knowledge.

Introduction

Hospitals generate a lot of data every day, including diagnoses, measurements, prescriptions, procedures, and more, which are stored in electronic health records (EHR). The adoption of EHR has greatly increased in many countries. The rate of EHR adoption now exceeds 96%, 94%, 88%, and 96% in the US, UK, China, and South Korea (hereafter Korea), respectively.[1-3] The widespread use of EHR has increased the amount of healthcare data, and numerous studies have been conducted using those data for various medical tasks, such as disease prediction, patient status monitoring, and diagnosis support.[4-6] For multicenter study, early works tried to use data from multiple hospitals; however, it was challenging to combine different EHR databases from each hospital without losing information because each database has its own purpose, structure, and terminology.[7] To overcome this problem, an observational medical outcomes partnership (OMOP)-common data model (CDM), which enables the transformation of different databases to a standardized format, was developed.[8,9] Using OMOP-CDM (hereafter CDM), statistical models or artificial intelligence methodologies for EHR data analysis can be shared throughout hospitals, and the analysis results can be validated externally using data from separate populations.

With the availability of large patient data, deep learning-based EHR analysis models have been developed, and learning patient representation through pretraining has become an area ripe for exploration.[10-15] As the pretraining-finetuning paradigm has achieved tremendous success in natural language processing, several studies proposed EHR pretraining models that learn patient representations from the sequential sets of medical codes corresponding to sentences in natural language. For pretraining tasks, most of those prior studies utilized bidirectional encoder representation from transformers (BERT)[16], one of the most famous models for contextualizing sequential inputs, especially natural language. As original BERT masks some of the tokens (words) from the sequential inputs (sentences) and infers the right tokens for the masked parts, BERT-based EHR pretraining models mask some of the medical codes and infer proper medical codes using unmasked preceding and following medical history. Prior studies achieved great performance in predicting several diseases, such as pancreatic cancer[13], heart failure[14], and non-accidental trauma[10], when finetuned initially pretrained models compared to initially randomized (not pretrained) models.

EHR pretraining models exhibited great performance in various medical tasks, but no study has dealt with adverse drug event (ADE) prediction using EHR pretraining models. Tertiary hospitals have good-quality inpatient records since all medical events that occur in the hospital are recorded in the EHR database. Therefore, EHR data from tertiary hospitals is well suited for monitoring the condition of hospitalized patients, especially

the risk of ADE.

In this study, we exhibited the validity of EHR pretraining model in ADE prediction tasks using CDM-based EHR data, which includes various domains such as diagnosis, prescription, measurement, and procedure. For pretraining, we adopted the masked language model (MLM). Codes were randomly masked, and EHR pretraining models were trained to infer the masked tokens with preceding and following history along with age and date. In this process, we introduced domain embedding (DE) to provide information about the domain of the masked token and prevent the model from finding unnecessary codes from other domains (Fig. 1a). EHR pretraining models with DE outperformed all the other baselines in several representative ADE prediction tasks. For the qualitative analysis, we identified important features at cohort and patient levels and demonstrated that the results were consistent with background clinical knowledge.

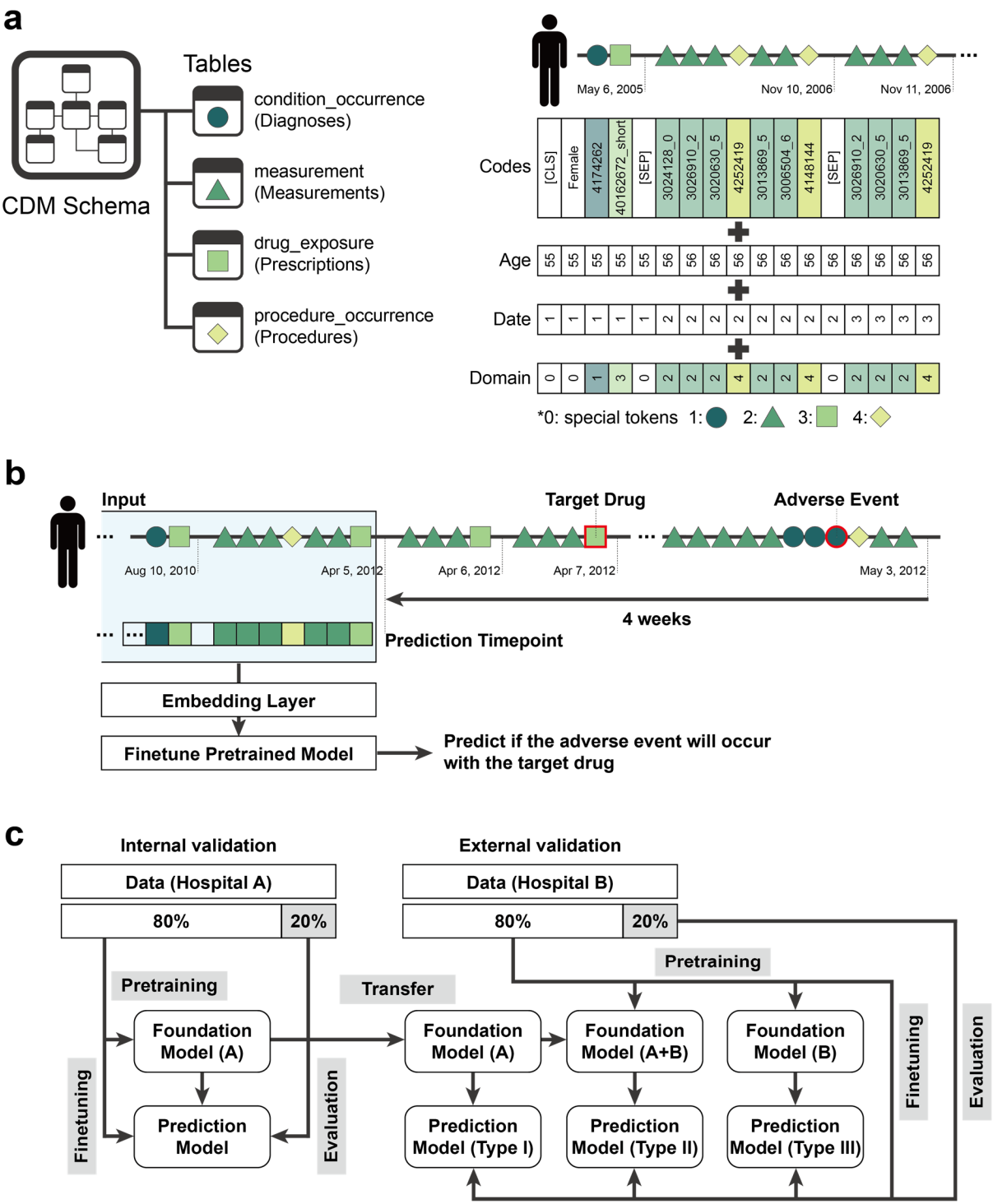

Fig. 1. Workflow of ADE prediction using CDM-based EHR pretraining model. a, Data from four domains (diagnosis, measurement, prescription, and procedure) of OMOP-CDM were utilized. Each patient trajectory consists of medical records sorted by time. All trajectories start with [CLS] token and gender token and are divided by [SEP] token for every day. Patient trajectory embedding is the summation of code, age, date, and domain embeddings. Domain embedding has five tokens corresponding to special tokens and four domains of CDM. **b**,

For the case group, ADE occurred within four weeks after prescription, and the prediction timepoint was four weeks before the occurrence of adverse events. Each patient trajectory before the prediction timepoint was utilized to predict the occurrence of ADE. **c**, For internal validation, we finetuned the pretrained foundation model to predict ADE and evaluated the prediction model. For external validation, we utilized three types of pretrained models: For type I, we transferred the foundation model pretrained by the internal dataset and finetuned the model with the external dataset. For type II, to mitigate the gap in data distribution between two separate hospitals, we used the external dataset to additionally pretrain the foundation model initially pretrained by the internal dataset. For type III, we pretrained the foundation model only using the external dataset.

Methods

Data curation

This study used data from the Seoul National University Hospital (SNUH) between January 2001 and December 2023 and the Ajou University Medical Center (AUMC) between January 2004 and December 2023. Both hospitals operate EHR database based on OMOP CDM version 5.3.[7] We collected data from four domains (tables) of CDM schema, diagnosis (condition_occurrence), prescription (drug_exposure), measurement (measurement), and procedure (procedure_occurrence). Terminologies of diagnosis and procedure, measurement, and prescription are based on SNOMED CT[17], Rx-Norm[18], and LOINC[19], respectively.

To screen patients with high-quality records, we included patients who had been hospitalized for at least three days in the study population, and patients under 18 were excluded. Data were randomly split into training (70%), validation (10%), and hold-out test (20%) datasets (Fig. 1c).

The Institutional Review Board (IRB) of Seoul National University Hospital (IRB approval No. 2204-001-1310) approved the study with a waiver of informed consent, considering that our study used retrospective and observational EHR data. The approval aligns with the principles outlined in the Declaration of Helsinki, the Korean Bioethics and Safety Act (Law No. 16372), and the Human Research Protection Program–Standard Operating Procedure of Seoul National University Hospital.

ADE prediction task

For drug groups, we included nonsteroidal anti-inflammatory drugs (NSAID), anticoagulants (AC), and chemotherapy (Chemo). For corresponding adverse events, we selected peptic ulcer (PU), intracranial hemorrhage (ICH), and neutropenic fever (NF), respectively. For each cohort, we first included patients who were prescribed the target drug. Those who were diagnosed with corresponding adverse events within four weeks after the prescription were included in the case group (Fig. 1b), and those who had no record of adverse events were included in the control group. For the case group, we set a prediction timepoint to four weeks before the adverse event occurrence. For the control group, we randomly selected one of the days on which the target drug was prescribed and set it as the prediction timepoint to minimize potential bias caused by selection criteria. Given that adverse events can vary in timing for different drugs and circumstances, we conducted additional experiments by setting the prediction timepoints to two, eight, and twelve weeks before the occurrence of adverse events.

Record tokenization

Each diagnosis or procedure record was tokenized by its corresponding concept ID in SNOMED-CT. For each prescription, we added 'short' at the end of the corresponding concept ID in Rx-Norm if its duration was shorter than four weeks and 'long' if it was more than four weeks (Fig. 1a). This process was to distinguish prescription type, considering inpatients are usually prescribed daily (or with every meal), while outpatients typically receive prescriptions for longer periods. Each measurement item was categorized into ten tokens divided into deciles, and we added a corresponding decile number between 0 and 9 at the end of its corresponding concept ID in LOINC (Fig. 1a). For example, in Fig. 1a, '40162672_short' indicates prescription of amitriptyline hydrochloride 10mg oral tablet with less than four weeks, and '3026910_2' indicates low level (between 2nd and 3rd decile) of gamma-glutamyl transferase (GGT). We only utilized numerical measurement items because the items with natural language results were unstructured and unformatted to tokenize. Regarding special tokens, we utilized five special tokens: [PAD], [MASK], [UNK], [CLS], and [SEP].[16] Like BERT used [SEP] to separate two sentences, we inserted [SEP] between all different dates. The token counts for each domain in each dataset are summarized in Supplementary Table 3. Ages were simply inserted as integer tokens, and dates were grouped by integer tokens increasing by one. We used five tokens for domains: special tokens, diagnosis, measurement, prescription, and procedure (Fig. 1a).

Trajectory construction and embedding

All tokens were sorted in order of time for each patient, and the maximum trajectory length was set to 2048, covering more than 85% of all trajectories. All trajectories start with [CLS] and gender tokens. For pretraining, trajectories that were longer than the maximum length were sliced into non-overlapping sub-trajectories with the maximum length to prevent potential dependency among trajectories from a single patient. For finetuning, we first removed tokens before prediction timepoint and used the latest 2048 tokens instead of slicing them into sub-trajectories. Trajectories with less than 50 tokens were excluded because predicting ADE in patients with too little information was not considered to be worthwhile. All short trajectories were padded to a length of 2048, aligning them with the same length for mini-batch training. For each token embedding, corresponding embeddings of age, date, and domain tokens were added.

Pretraining process

EHR pretraining models were based on BERT[13,16] architecture, with six layers, eight attention heads, and a hidden dimension of 256. We aggregated the vocabulary sets from the internal and external datasets, and the final vocabulary size was 41,536. However, we set the embedding size of vocabulary to be 50,000, considering future additional medical codes. For age, date, and domain embedding, we set the embedding size to 180, 1024, and 20, respectively. The size of domain embedding was decided considering additional domains like unstructured data (e.g., X-ray, electrocardiogram, and nursing report). For pretraining, we randomly masked 30% of tokens except special tokens for each trajectory, and the model was trained to infer the masked tokens of specific domains with preceding and following history along with age, date, and domain. The models were pretrained with 100 epochs, and we selected the model with the minimum CE loss during the training.

For external validation, we introduced three types of EHR pretraining model (Fig. 1c): For type I, we used the model pretrained with the internal dataset without additional pretraining. For type II, we used the external dataset to additionally pretrain the model initially pretrained by the internal dataset. For type III, we initially pretrained the initial model using the external dataset.

Finetuning process

For finetuning for ADE prediction, we added feed-forward neural network (FFNN), hyperbolic tangent (Tanh) activation function, dropout, and FFNN on the representation of the first token to yield a vector of size two for binary classification. Finetuning was trained to minimize binary CE loss for ADE prediction. We recorded the classification performance for every cycle of 100 batches, and the finetuning was stopped if AUROC of the validation dataset did not improve for ten cycles (1000 batches). Considering extreme class imbalance and small batch size, we randomly oversampled the minority class to be one-tenth of the majority class during finetuning.

Feature importance

We calculated the importance of features using attention matrices of model outputs in case groups. We averaged attention matrices of eight attention heads from the last self-attention layer to yield a final attention matrix (hereafter attention matrix). An attention matrix A for each trajectory is 2048 by 2048 matrix (all trajectories are aligned to have 2048 tokens), and an attention score $\mathrm{A_j}$ indicates how closely $\mathrm{i^{th}}$ and $\mathrm{j^{th}}$ tokens are related.[20] Note that $\sum_{\mathrm{j}} A_{ij} = 1 \; \forall \mathrm{i} \in (0, \ldots, 2048)$.

For comprehensive analysis, we first averaged the values of the same tokens in each row of the attention

matrix to prevent repeatedly appearing tokens of routine medical events (e.g., normal saline, blood pressure, and electrocardiogram) from getting unnecessarily high scores. Then, we extracted the maximum value among 2048 rows of each token. Finally, each 2048 by 2048 attention matrix was transformed into a one-dimensional vector. We called this vector as trajectory attention vector (TAV). The value of each element of TAV is as follows:

$$\mathrm{TAV}_{\mathrm{i_v}} = \max_{i} \frac{\sum_{j \in I_{A,v}} A_{ij}}{|I_{A,v}|}$$

where i_v is the index of code v in TAV and $I_{A,v}$ is a set of indices of code v in the original attention matrix A. TAV finally indicates the importance of each token in the trajectory. In each TAV, we excluded tokens that appeared in fewer than 5% of the patients in the case group. This process was to prevent irrelevant trivial tokens coincidently having high attention scores from becoming important features. Finally, we averaged all TAV values by tokens and selected the top 10 most important tokens from each domain.

For individualized analysis, we simply summed all rows of an attention matrix and then extracted the top 10% of tokens with the highest values. This process was to enhance readability because most trajectories had hundreds of tokens.

Statistical analysis

Characteristics (age, sex, and comorbidities) in internal and external cohorts were compared by calculating P-values using the Student's t-test for age and the Fisher's exact test for the rest of the variables. To measure and compare the performances of the models, we mainly used AUROC. We also provided area under the precision-recall curve (AUPRC), sensitivity, specificity, precision, and F1-score. Confidence intervals of AUROC and AUPRC were calculated by DeLong's method[21], while those of F1-score, sensitivity, specificity, and precision were calculated by Wilson's method[22]. Statistical significance was set at α=0.05. All statistical analyses were performed using scikit-learn (version 1.0.2) in Python (version 3.8.10).

Training and computing information

The learning rate and dropout rate were set to 5e-5 and 0.1, and the batch size was set to 16. For deep learning and BERT implementation, we used Pytorch (version 1.12.0) and HuggingFace package (version 4.41.2)[23] in Python (version 3.8.10). We used four NVIDIA RTX A6000 GPUs for internal processes and four NVIDIA GeForce RTX 3090 Ti GPUs for external processes.

Results

Study population

We used records of 510,879 and 419,505 adult patients (aged over 18) who were hospitalized for at least three days in SNUH and AUMC, respectively. The development dataset was used to pretrain the model and to finetune the pretrained model for ADE prediction. The hold-out test datasets were used to evaluate the performance of ADE prediction models. The baseline characteristics of the datasets are summarized in Table 1. The vocabulary size was higher in the internal dataset (SNUH), while the median number of medical codes per patient was higher in the external dataset (AUMC). We aggregated the vocabulary sets of the two hospitals' development datasets, and the size of the final vocabulary was 41,536, with around 7,000 codes overlapped. Even though those two hospitals share the same data structure and vocabulary system, many mapped codes were different. The distribution of comorbidities was also different between the two hospitals. The rates of dementia, renal disease, and malignant tumor were higher in the internal dataset, while the rest of the comorbidities were higher in the external dataset.

Table 1. Baseline characteristics of datasets

	Internal dataset (SNUH)	External dataset (AUMC)
The number of patients	510,879	419,505
Mean age (SD)	55.09 (17.19)	53.84 (18.75)
Male sex	47.10%	50.51%
Median days of hospitalization (IQR)	28 (15, 63)	30 (14, 71)
Median number of codes per patient (IQR)	476 (228, 1130)	627 (283, 1545)
Vocabulary size	26,417	22,551
Comorbidities		
Myocardial infarction	1.18%	2.57%
Congestive heart failure	2.04%	2.98%
Peripheral vascular disease	1.28%	1.83%
Cerebrovascular disease	7.57%	9.81%
Dementia	4.98%	3.46%
Pulmonary disease	5.86%	8.51%
Connective tissue disease	1.63%	4.34%
Peptic ulcer disease	3.80%	6.03%
Mild liver disease	5.25%	8.63%
Liver disease	0.70%	1.71%
Uncomplicated diabetes	8.10%	14.36%
Complicated diabetes	1.50%	4.64%
Paraplegia and hemiplegia	0.50%	1.70%
Renal disease	4.05%	4.02%
Malignant tumor	23.70%	18.37%
Metastatic carcinoma	0.82%	1.11%

SNUH, Seoul National University Hospital; AUMC, Ajou University Medical Center; SD, standard deviation; IQR, interquartile range

Patient demographics of each cohort from both hospitals are summarized in Supplementary Table 4. As the prediction timepoint moved back to the past and the monitoring period for adverse events increased, the number of patients in each case group also increased in NSAID-PU and AC-ICH cohorts. For the Chemo-NF cohort, more patients were excluded due to a lack of previous records than those included by an increased monitoring period.

Pretraining results

We adopted DE to prevent the model from finding unnecessary medical codes from other domains during the inference of masked tokens in the pretraining process, the validation losses of MLMs were much lower with DE (Supplementary Fig. 1). This result indicates that finding the correct codes for masked tokens was much easier with a hint of domain. For pretraining external validation, The validation loss of type II pretrained models was much lower than the type III, even though those models were pretrained by the same external dataset. In addition, even with a few epochs, the validation loss of the type II models showed convergence (Supplementary Fig. 1b). It took around a week to train 100 epochs for both models with and without DE for both internal and external datasets.

Performance evaluation

We evaluated ADE prediction performance with four models: Initially randomized models with and without DE, the model pretrained with and without DE. For internal validation, we pretrained the initial model (with and without DE) for 100 epochs and finetuned the model using each cohort dataset. During the finetuning process, the area under the receiver operating characteristic curve (AUROC) for prediction was calculated for every 100 batches, and the finetuning process was stopped if the AUROC did not increase for ten cycles (1000 batches). The model pretrained with DE outperformed all the other models in all metrics in all cohorts (Table 2). The pretraining process improved AUROC and AUPRC for all cohorts (pretrained without DE), but the adoption of DE improved the performance much more. When the model was initialized with randomized weights, the adoption of DE hardly influenced the performance.

Table 2. Internal validation of finetuned models for all cohorts.

	AUROC	AUPRC	Sensitivity	Specificity	Precision	F1-score
NSAID (PU)						
Randomized without DE	0.848 (0.833-0.864)	0.044 (0.029-0.060)	0.836 (0.832-0.839)	0.727 (0.724-0.731)	0.025 (0.024-0.027)	0.049 (0.047-0.051)
Randomized with DE	0.848 (0.833-0.864)	0.044 (0.029-0.060)	0.836 (0.832-0.839)	0.727 (0.723-0.731)	0.025 (0.024-0.027)	0.049 (0.047-0.051)
Pretrained without DE	0.954 (0.946-0.962)	0.357 (0.349-0.365)	0.909 (0.906-0.911)	0.870 (0.867-0.873)	0.056 (0.054-0.058)	0.105 (0.103-0.108)
Pretrained with DE	**0.977 (0.971-0.984)**	**0.668 (0.661-0.675)**	**0.942 (0.940-0.944)**	**0.916 (0.913-0.918)**	**0.087 (0.084-0.089)**	**0.159 (0.156-0.162)**
AC (ICH)						
Randomized without DE	0.769 (0.739-0.799)	0.022 (0.000-0.799)	0.753 (0.749-0.758)	0.683 (0.678-0.688)	0.016 (0.015-0.018)	0.032 (0.030-0.034)
Randomized with DE	0.769 (0.739-0.799)	0.022 (0.000-0.799)	0.753 (0.749-0.758)	0.683 (0.678-0.688)	0.016 (0.015-0.018)	0.032 (0.030-0.034)
Pretrained without DE	0.822 (0.797-0.848)	0.040 (0.015-0.065)	0.836 (0.831-0.840)	0.670 (0.665-0.675)	0.017 (0.016-0.019)	0.034 (0.032-0.036)
Pretrained with DE	**0.908 (0.886-0.931)**	**0.206 (0.183-0.228)**	**0.840 (0.836-0.844)**	**0.863 (0.859-0.867)**	**0.041 (0.039-0.043)**	**0.078 (0.075-0.081)**
Chemo (NF)						
Randomized without DE	0.833 (0.817-0.848)	0.140 (0.124-0.156)	0.867 (0.862-0.872)	0.669 (0.662-0.676)	0.072 (0.068-0.076)	0.132 (0.127-0.137)
Randomized with DE	0.833 (0.817-0.849)	0.140 (0.124-0.156)	0.867 (0.862-0.872)	0.668 (0.661-0.675)	0.071 (0.068-0.075)	0.132 (0.127-0.137)
Pretrained without DE	0.971 (0.966-0.976)	0.704 (0.698-0.709)	0.933 (0.929-0.936)	0.890 (0.885-0.895)	0.200 (0.194-0.206)	0.329 (0.322-0.337)
Pretrained with DE	**0.989 (0.985-0.993)**	**0.909 (0.905-0.912)**	**0.953 (0.950-0.956)**	**0.948 (0.945-0.951)**	**0.350 (0.343-0.358)**	**0.512 (0.505-0.520)**

Bold indicates the best. Sensitivity, specificity, precision, and F1-score were calculated by Youden's index. Confidence intervals (CIs) of AUROC and AUPRC were calculated by DeLong's method. CIs of sensitivity, specificity, precision, and F1-score were calculated by Wilson's method. NSAID, nonsteroidal anti-inflammatory drugs; PU, peptic ulcer; AC, anticoagulants; ICH, intracranial hemorrhage; Chemo, chemotherapy; NF, neutropenic fever; AUROC, area under the receiver operating characteristic curve; AUPRC, area under the precision-recall curve

For external validation, the type II and III models were pretrained for ten epochs. The adoption of DE improved AUROC and AUPRC of the type I and II models in all cohorts (Table 3). For the type I with DE, even though the model was pretrained from the different hospital data (SNUH) and there was no additional pretraining, the AUROC was much higher than the randomized model in all cohorts (AUROC from 0.814 to 0.947 in NSAID-PU cohort, from 0.769 to 0.940 in AC-ICH cohort, and from 0.918 to 0.934 in Chemo-NF cohort). For the type II and III, the type II exhibited higher AUROC and AUPRC in all cohorts. Considering those two models were pretrained for the same ten epochs, the model initially pretrained by the other dataset (type II) was more effective than the initially randomized model (type III). Like internal validation results, the adoption of DE had minimal effect on performance in initially randomized models. The receiver operating characteristic (ROC) curves and precision-recall (PR) curves of internal and external validations are summarized in Fig. 2. Since the proportion of patients with adverse events was much less than those without (Supplementary Table 3), most models overpredicted the risk of adverse events (Fig. 3).

Table 3. External validation of finetuned models for all cohorts.

	AUROC	AUPRC	Sensitivity	Specificity	Precision	F1-score
NSAID (PU)						
Randomized without DE	0.814 (0.798-0.830)	0.070 (0.054-0.086)	0.804 (0.800-0.808)	0.691 (0.686-0.696)	0.042 (0.040-0.044)	0.079 (0.076-0.082)
Randomized with DE	0.814 (0.798-0.830)	0.070 (0.054-0.086)	0.804 (0.800-0.808)	0.690 (0.685-0.695)	0.041 (0.039-0.044)	0.079 (0.076-0.082)
Pretrained without DE (Type I)	0.834 (0.819-0.848)	0.077 (0.063-0.092)	0.793 (0.788-0.797)	0.724 (0.719-0.729)	0.046 (0.044-0.048)	0.086 (0.083-0.089)
Pretrained with DE (Type I)	0.947 (0.938-0.956)	0.575 (0.566-0.584)	0.847 (0.843-0.851)	0.904 (0.901-0.907)	0.128 (0.124-0.131)	0.222 (0.218-0.226)
Pretrained without DE (Type II)	0.919 (0.908-0.929)	0.278 (0.268-0.289)	0.867 (0.864-0.871)	0.826 (0.822-0.830)	0.077 (0.074-0.079)	0.141 (0.137-0.144)
Pretrained with DE (Type II)	**0.971 (0.964-0.978)**	**0.712 (0.705-0.719)**	**0.908 (0.905-0.911)**	**0.930 (0.927-0.933)**	**0.178 (0.174-0.182)**	**0.297 (0.292-0.302)**
Pretrained without DE (Type III)	0.857 (0.844-0.871)	0.118 (0.104-0.132)	0.883 (0.879-0.886)	0.679 (0.674-0.684)	0.044 (0.042-0.046)	0.083 (0.081-0.086)
Pretrained with DE (Type III)	0.925 (0.913-0.936)	0.435 (0.423-0.446)	0.857 (0.853-0.861)	0.839 (0.836-0.843)	0.082 (0.079-0.084)	0.149 (0.145-0.153)
AC (ICH)						
Randomized without DE	0.769 (0.738-0.799)	0.043 (0.013-0.074)	0.735 (0.729-0.741)	0.678 (0.672-0.684)	0.022 (0.020-0.024)	0.043 (0.040-0.045)
Randomized with DE	0.769 (0.738-0.799)	0.044 (0.013-0.074)	0.735 (0.729-0.741)	0.678 (0.672-0.683)	0.022 (0.020-0.024)	0.043 (0.040-0.045)
Pretrained without DE (Type I)	0.856 (0.831-0.881)	0.265 (0.240-0.290)	0.774 (0.768-0.779)	0.790 (0.785-0.795)	0.035 (0.033-0.037)	0.067 (0.064-0.070)
Pretrained with DE (Type I)	0.940 (0.923-0.958)	0.570 (0.552-0.588)	0.855 (0.850-0.859)	0.891 (0.887-0.895)	0.072 (0.068-0.075)	0.132 (0.128-0.137)
Pretrained without DE (Type II)	0.943 (0.927-0.959)	0.573 (0.557-0.589)	0.838 (0.833-0.842)	**0.900 (0.897-0.904)**	**0.076 (0.073-0.080)**	**0.140 (0.136-0.144)**
Pretrained with DE (Type II)	**0.972 (0.963-0.981)**	**0.673 (0.664-0.681)**	**0.957 (0.955-0.960)**	0.853 (0.849-0.858)	0.060 (0.057-0.063)	0.113 (0.109-0.117)
Pretrained without DE (Type III)	0.921 (0.902-0.940)	0.469 (0.450-0.487)	0.838 (0.833-0.842)	0.865 (0.861-0.870)	0.058 (0.055-0.061)	0.108 (0.104-0.112)
Pretrained with DE (Type III)	0.909 (0.887-0.930)	0.532 (0.511-0.554)	0.791 (0.785-0.796)	0.869 (0.864-0.873)	0.056 (0.053-0.059)	0.104 (0.101-0.108)
Chemo (NF)						
Randomized without DE	0.918 (0.893-0.942)	0.236 (0.212-0.261)	0.892 (0.884-0.899)	0.805 (0.795-0.814)	0.053 (0.048-0.058)	0.099 (0.093-0.107)
Randomized with DE	0.918 (0.893-0.942)	0.236 (0.212-0.261)	0.892 (0.884-0.899)	0.805 (0.795-0.814)	0.053 (0.048-0.058)	0.099 (0.093-0.107)
Pretrained without DE (Type I)	0.909 (0.885-0.933)	0.117 (0.093-0.141)	0.904 (0.896-0.910)	0.788 (0.779-0.798)	0.049 (0.045-0.055)	0.094 (0.087-0.101)
Pretrained with DE (Type I)	0.934 (0.915-0.953)	0.210 (0.191-0.229)	0.928 (0.921-0.934)	0.832 (0.823-0.841)	0.063 (0.058-0.069)	0.118 (0.111-0.126)
Pretrained without DE (Type II)	0.938 (0.921-0.955)	0.237 (0.220-0.254)	0.880 (0.872-0.887)	0.864 (0.855-0.872)	**0.073 (0.067-0.079)**	**0.135 (0.127-0.143)**
Pretrained with DE (Type II)	**0.951 (0.937-0.966)**	**0.368 (0.354-0.383)**	**0.988 (0.985-0.990)**	0.768 (0.758-0.778)	0.049 (0.044-0.055)	0.094 (0.087-0.101)
Pretrained without DE (Type III)	0.933 (0.910-0.955)	0.278 (0.255-0.300)	0.940 (0.934-0.945)	0.798 (0.789-0.808)	0.054 (0.049-0.059)	0.102 (0.095-0.109)
Pretrained with DE (Type III)	0.925 (0.902-0.948)	0.231 (0.208-0.254)	0.831 (0.822-0.840)	**0.867 (0.859-0.875)**	0.071 (0.065-0.077)	0.130 (0.123-0.138)

Bold indicates the best. Sensitivity, specificity, precision, and F1-score were calculated using Youden's index. Confidence intervals (CIs) of AUROC and AUPRC were calculated using DeLong's method. CIs of sensitivity, specificity, precision, and F1-score were calculated using Wilson's method. Type I indicates the foundation model initially pretrained by the internal dataset. Type II indicates the foundation model that was initially pretrained by the internal dataset and was additionally pretrained by the external dataset. Type III indicates the foundation model that was pretrained only by the external dataset. NSAID, nonsteroidal anti-inflammatory drug; PU, peptic ulcer; AC, anticoagulant; ICH, intracranial hemorrhage; Chemo, chemotherapy; NF, neutropenic fever; AUROC, area under the receiver operating characteristic curve; AUPRC, area under the precision-recall curve

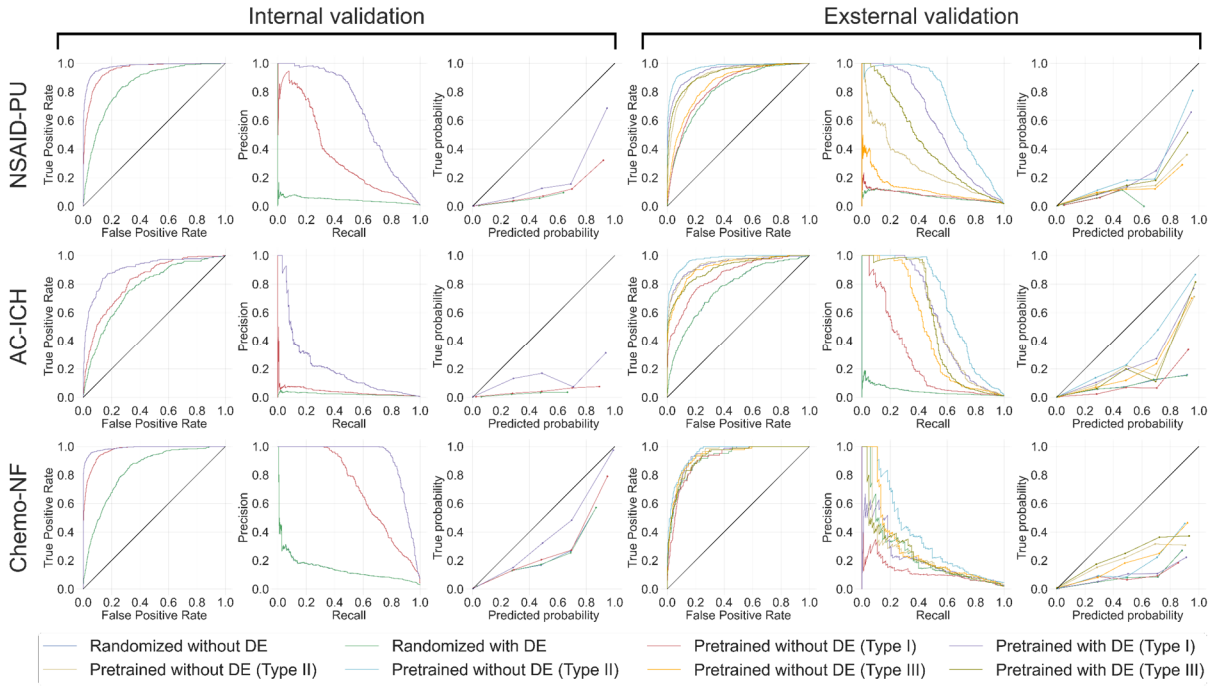

Fig. 2. Receiver operating characteristic (ROC), precision-recall (PR), and calibration curves of internal and external validations. The left three columns are the results of internal validation and the right three columns are the results of external validation. The prediction cutoff point of each finetuned model was set at Youden's index, which maximizes the sum of sensitivity (true positive rate) and specificity (1 - false positive rate). For both internal and external validation, most of the blue lines (for models randomized without DE) are invisible because they were overlapped by the green lines (for models randomized with DE).

Analysis of different prediction timepoints

We set the prediction timepoint to four weeks before the occurrence of adverse events. However, since adverse events can occur at different times depending on the drug and circumstances, we conducted additional experiments by setting the prediction timepoints to two, eight, and twelve weeks before the occurrence of adverse events. For external validation, we utilized the type II pretrained model with DE because it was the best-performing type with a prediction timepoint of four weeks before ADE occurrence (Table 3). The model pretrained with DE exhibited the best AUROC and AUPRC in all cohorts with all timepoints, only except the Chemo-NF cohort of the external dataset with the prediction timepoint of twelve weeks (Supplementary Tables 5-10). Adoption of DE was effective in most cases with different prediction timepoints.

Model interpretation

We reported the top 10 most important features from each domain, and various features relevant to each drug and adverse event were included in the top 10 (Figs. 3 and 4). In this process, we used the results of the model pretrained with DE (type II for external validation). For the NSAID-PU cohort in the internal validation, the model mainly focused on NSAIDs and aspirin prescription patterns for predicting PU. Celecoxib was the most important, even though it is known to be associated with a lower risk of gastrointestinal adverse events (such as PU).[24,25] This might be due to patients switching to celecoxib after experiencing gastrointestinal symptoms while taking other NSAIDs; indeed, patients who took celecoxib were prescribed an average of 3.1 NSAIDs, while those not taking celecoxib were prescribed an average of 1.6 NSAIDs. Clopidogrel is usually prescribed with aspirin for the prevention of cardiovascular events.[26] Clopidogrel can be associated with several diagnoses and procedures regarding cardiovascular diseases (angina pectoris, hypertension, cerebral infarction, and all of the included procedures). Gastric symptoms such as chest pain and nonulcer dyspepsia were also important. In the external validation, similarly, the model paid attention to NSAIDs and aspirin prescription patterns. For diagnosis, several conditions that are usually treated by NSAIDs (headache, joint pain, and pain caused by periodontitis) were important, and like the internal validation, chest pain was one of the main factors. Headache can be associated with another important feature, cerebral infarction. Valid medical context existed; however, more investigation is needed into the reason why cardiovascular diseases were important in the NSAID-PU cohort.

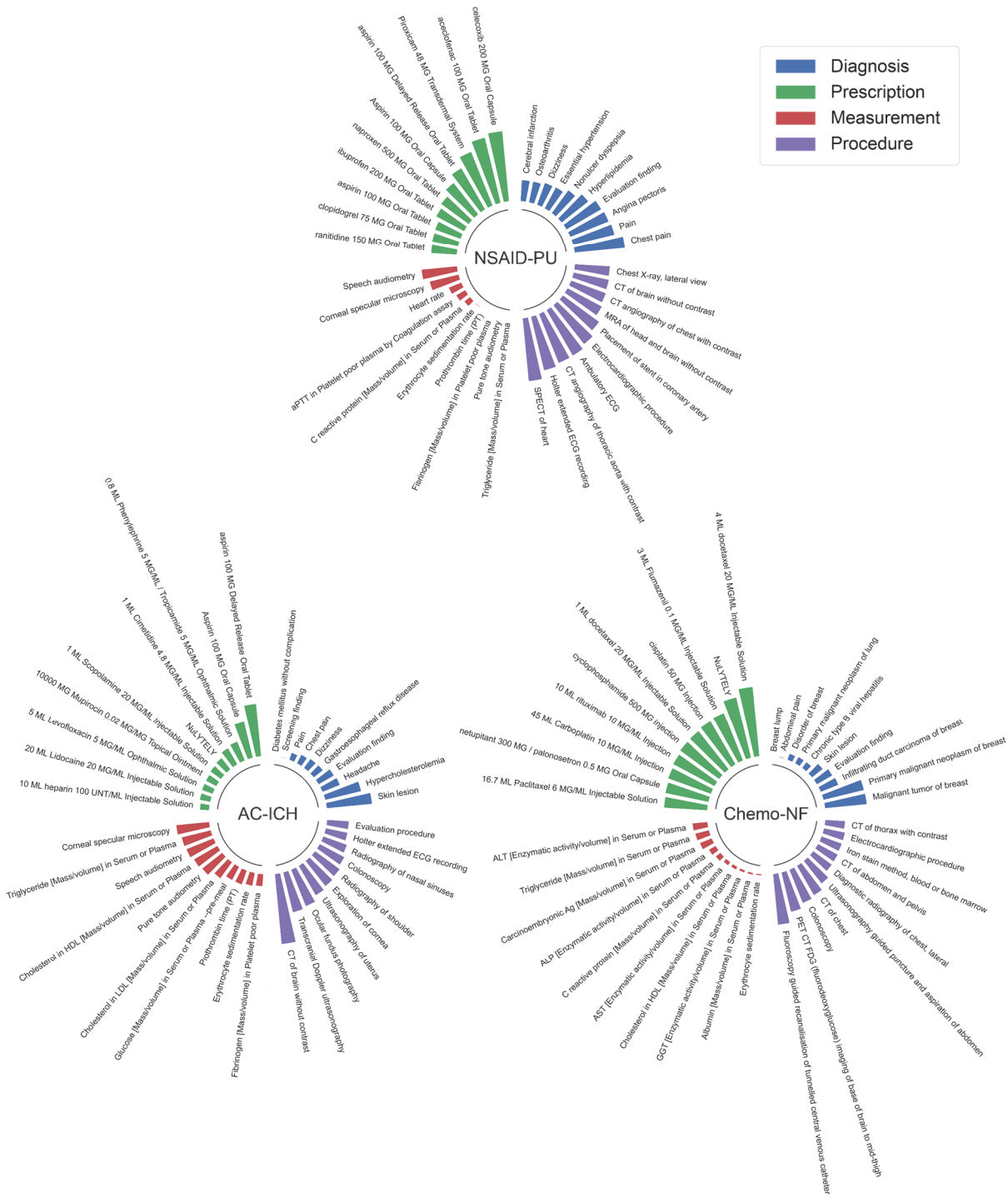

Fig. 3. Feature importance of models finetuned with the internal dataset. Top 10 important features in each domain. Feature importance was deduced based on attention scores in the model. To prevent routine medical events (e.g., normal saline, blood pressure, and electrocardiogram) that repeatedly appear in trajectories to be excessively important, attention scores of the same tokens were initially averaged at the trajectory level, and then the scores of each token were finally averaged. For each cohort, the length of the longest bar was fixed to be the same, and the lengths of all the rest of the features were normalized accordingly.

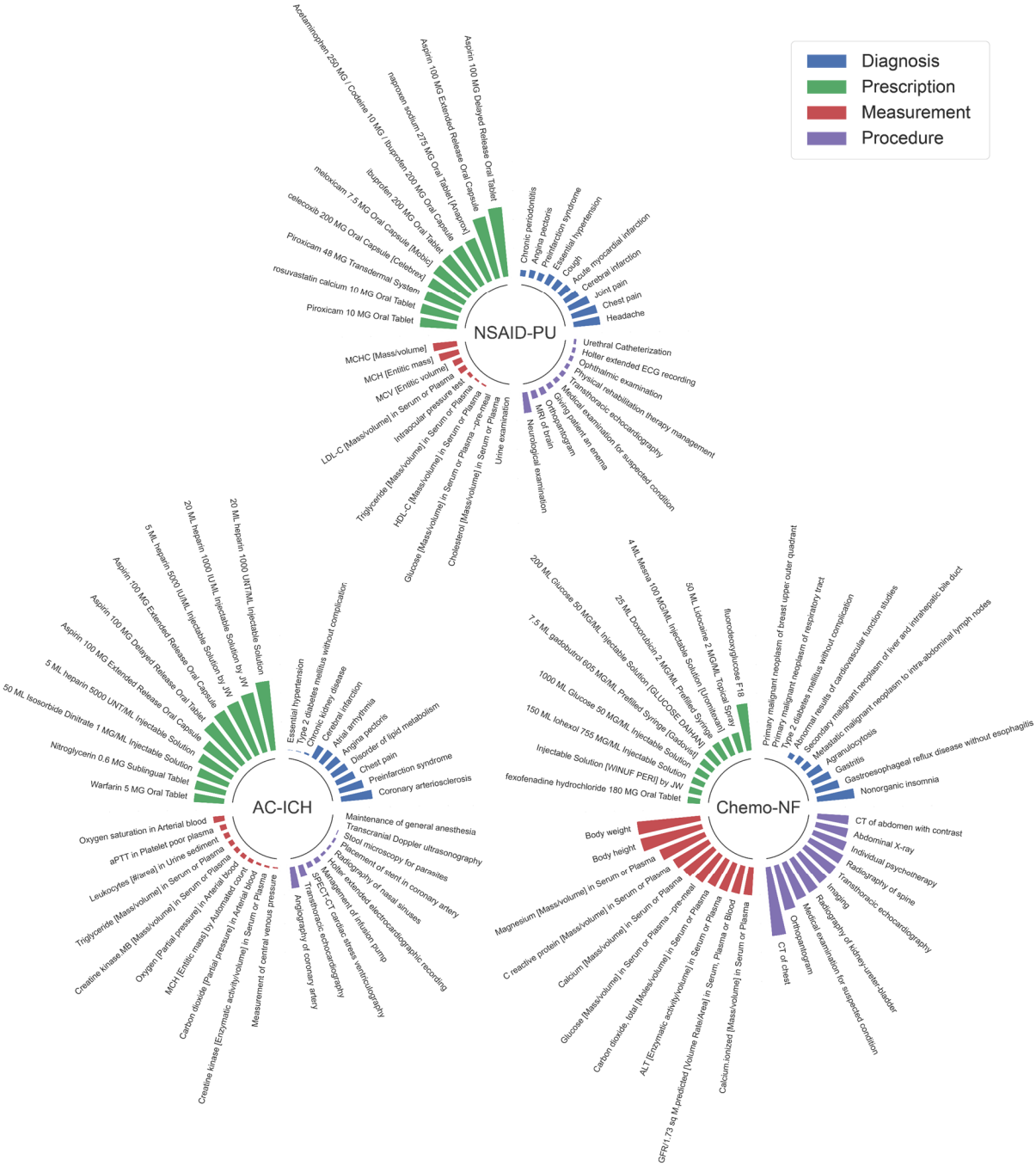

Fig. 4. Feature importance of models finetuned with the external dataset. Top 10 important features in each domain. The feature importance calculation process was the same as that used for the internal dataset.

For the AC-ICH cohort in the internal validation, the models perceived aspirin and heparin prescription patterns as important for ICH prediction. The most important diagnosis was skin lesion, which might be related to bruising or petechiae caused by AC, and clinicians might have simply coded the condition as a skin lesion rather than its own name.[27,28] Headache and dizziness, which are the main symptoms of ICH, were also important.[29,30]

The second most important diagnosis, hypercholesterolemia, is closely related to cardiovascular disease and can be associated with ICH.[31,32] For measurements, lipid profile (triglyceride, HDL-cholesterol, LDL-cholesterol), which is a direct indicator for assessing hypercholesterolemia, and prothrombin time (PT) and fibrinogen, which are primarily focused on ICH patients,[33-35] were included in the top 10 most important measurements. The models also closely observed procedures for diagnosis of cerebrovascular diseases, such as computed tomography (CT) of brain and transcranial Doppler (TCD) ultrasonography. The feature importance calculation result of the internal validation well-reflected prior knowledge regarding AC and ICH, but an unexpected feature (gastroesophageal reflux disease (GERD)) still existed. In the external validation, there were many features regarding heart diseases: coronary arteriosclerosis, preinfarction syndrome, angiography of coronary artery, and transthoracic echocardiography. Compared to the internal validation, the model paid less attention to the measurement, but several factors associated with heart disease (aPTT and creatine kinase-MB)[36,37] were included in the top 10. Accordingly, in this cohort, the patients with heart disease might have developed cerebrovascular disease as well.

For the CT-NF cohort in the internal validation, several drugs for chemotherapy (docetaxel, cisplatin, cyclophosphamide, rituximab, carboplatin, and paclitaxel) were included. Netupitant and palonosetron are drugs for preventing chemotherapy-induced nausea and vomiting.[38,39] Patients who had experienced breast cancer were more likely to suffer from NF in this cohort. Chronic type B viral hepatitis is closely related to hepatocellular carcinoma. For measurement, carcinoembryonic antigen, which is an indicator for cancer diagnosis and monitoring,[40] and inflammation-related features such as C reactive protein (CRP) and erythrocyte sedimentation rate (ESR) were included in the top 10 most important measurements. Central venous catheter (CVC) and PET CT FDG, which are closely associated with cancer, were important procedures. Screening colonoscopy is essential for the detection of colon cancer, and Nulytely is used for bowel cleansing prior to colonoscopy.[41,42] Several types of CT and ultrasonography guided puncture and spiration of abdomen, which are procedures for the detection of cancer, were also included. Iron stain method, which is usually executed for the detection of hematologic malignancy, one of the highest-risk cancers for NF, was another important procedure.[43,44] In the external validation, fluorodeoxyglucose (FDG) F18, gadobutrol, and iohexol, which are contrast medium drugs for cancer detection, were focused on more than chemotherapy drugs. Contrary to the internal validation, it is notable that agranulocytosis, which usually precedes NF,[45] was included in the top 10.

In addition to cohort-level interpretation, patient-level interpretation is also available. We reported a sample of patient-level interpretation in the Chemo-NF cohort (Supplementary Fig. 2). Patient-level interpretation allows clinicians to understand which records and timepoints were important for predicting ADE for each patient.

Discussion

We have proposed, externally validated, and qualitatively analyzed EHR pretraining model pretrained on CDM-based EHR data with DE. CDM schema consists of several domains, and we utilized data from four domains, including diagnosis, measurement, prescription, and procedure. We adopted DE in EHR pretraining model to let the model focus only on a necessary domain, and DE improved the performance of almost all the ADE prediction tasks. For external validation, the type II pretrained model outperformed all the other models, demonstrating the effectiveness of the model initially pretrained by a different dataset. The qualitative analysis with the importance of features at the cohort and patient levels was available, and the attention scores of each finetuned model well-reflected the relevant background clinical knowledge and the characteristics of the corresponding cohort.

We validated that the initially pretrained model works with another dataset. In the external validation, the type I pretrained model outperformed initially randomized models in all cohorts, even though they were never aware of the external dataset. In addition, the type II pretrained model achieved significantly improved performance with only ten training epochs. Given that only 7,000 codes were common between the two hospitals, which had 26,417 and 22,551 codes, respectively, it is noteworthy that the model pretrained with numerous unknown codes remained effective. Additional pretraining successfully resolved the disparity of different datasets and improved performance. The initially pretrained model was also effective in terms of time efficiency. It took around six hours to pretrain the external dataset for one iteration with our single graphic processing unit (GPU). Considering we spent around a week for 100 iterations with four multiple GPUs, the ability to develop a strong pretrained model with fewer epochs can be especially attractive to smaller institutions with limited computational resources.

There are several studies that dealt with the EHR pretraining model, but this is the first study that dealt with ADE. Hospital data has clear pros and cons for EHR pretraining: The upside is that the patient data generated within the hospital is real-time and of very high quality, and the downside is that it is very difficult to access data outside the hospital. Survey data can supplement historical medical records to some extent, but it is not enough. Accordingly, it is difficult to predict diseases, mortality, and readmission, which are very affected by data outside the hospital, using in-hospital data.[12-15,46] To maximize the advantages of hospital data, we applied the EHR pretraining model to ADE prediction for inpatients. A pretraining model can learn changes before and after taking medication. Thus, the pretraining process included a deep understanding of drug reactions.

Compatibility is a significant advantage of CDM-based EHR pretraining model. OMOP CDM is an open community data standard and is represented in more than 19 countries, with more than 200 million patient

records, and more than 2,500 collaborators.[7,47] The model pretrained from one hospital can be applied to, enhanced by, and validated by many other global hospital datasets.[8,9] In addition, there is a national institution called Health Insurance Review and Assessment Service (HIRA) in Korea that has claims data converted to OMOP CDM format, and claims data can complement the lack of patient follow-up in hospital data.[48-50] A large scale model through federated learning algorithms is also available.[51] Several studies have already conducted multi-institutional data analysis via the federated learning framework for OMOP CDM data.[52,53] CDM-based EHR pretraining model trained by data at the regional or national level can serve as a foundation model for various tasks in addition to in-hospital tasks such as ADE prediction.

This study has several limitations. First, we defined cohorts very roughly, even though several critical exclusion criteria might exist for proper cohort analysis. Second, there were irrelevant important features for each cohort. For example, speech audiometry, which is extraneous with all the ADE in this study, was found in the NSAID-PU and AC-ICH cohorts. It might be associated with overfitting, black box, and age, as PU and ICH usually occur in aged patients, but further investigation is needed. Third, the percentage of shared medical codes between the internal and external datasets was too low. Although this discrepancy demonstrated the effectiveness of the model pretrained with numerous unknown codes, alternatives to mitigate the inconsistency of vocabulary should be investigated more. Fourth, our model was only trained by structured EHR data. For future work, utilizing several types of unstructured medical data and combining those data types are suggested.

In conclusion, we have demonstrated the potential of CDM-based EHR pretraining model as a foundation model through its prediction performance, interpretability, and compatibility. The adoption of domain embedding was effective in almost all cases, simplifying the pretraining procedures and improving comprehension of a medical context. The model interpretability for each cohort was consistent with several prior studies and clinical knowledge. Enhanced code systems to mitigate vocabulary inconsistency and CDM-based EHR pretraining model combined with unstructured data types are suggested for future works.

Data availability

The datasets of SNUH and AUMC are not publicly available due to patient privacy and are available from the corresponding author on reasonable request and IRB approvals.

Code availability

Source code for the experiments is publicly available at https://github.com/kicarussays/CDM-BERT.

References

1 Parasrampuria, S. & Henry, J. Hospitals' use of electronic health records data, 2015–2017. *ONC Data Brief* **46**, 13 (2019).

2 Liang, J. *et al.* Adoption of electronic health records (EHRs) in China during the past 10 years: consecutive survey data analysis and comparison of sino-american challenges and experiences. *Journal of medical Internet research* **23**, e24813 (2021). https://doi.org:10.2196/24813

3 Lee, K. *et al.* Digital Health Profile of South Korea: A Cross Sectional Study. *International Journal of Environmental Research and Public Health* **19**, 6329 (2022).

4 Li, Y., Li, Y. & Tian, H. Deep Learning-Based End-to-End Diagnosis System for Avascular Necrosis of Femoral Head. *IEEE Journal of Biomedical and Health Informatics* **25**, 2093-2102 (2021). https://doi.org:10.1109/JBHI.2020.3037079

5 Poongodi, T., Sumathi, D., Suresh, P. & Balusamy, B. in *Bio-inspired Neurocomputing* (eds Akash Kumar Bhoi, Pradeep Kumar Mallick, Chuan-Ming Liu, & Valentina E. Balas) 73-103 (Springer Singapore, 2021).

6 Ramkumar, G., Seetha, J., Priyadarshini, R., Gopila, M. & Saranya, G. IoT-based patient monitoring system for predicting heart disease using deep learning. *Measurement* **218**, 113235 (2023). https://doi.org:https://doi.org/10.1016/j.measurement.2023.113235

7 Biedermann, P. *et al.* Standardizing registry data to the OMOP Common Data Model: experience from three pulmonary hypertension databases. *BMC Medical Research Methodology* **21**, 238 (2021). https://doi.org:10.1186/s12874-021-01434-3

8 Stang, P. E. *et al.* Advancing the Science for Active Surveillance: Rationale and Design for the Observational Medical Outcomes Partnership. *Annals of Internal Medicine* **153**, 600-606 (2010). https://doi.org:10.7326/0003-4819-153-9-201011020-00010

9 Bathelt, F. The usage of OHDSI OMOP–a scoping review. *Proceedings of the German Medical Data Sciences (GMDS)*, 95 (2021).

10 Huang, D., Cogill, S., Hsia, R. Y., Yang, S. & Kim, D. Development and external validation of a pretrained deep learning model for the prediction of non-accidental trauma. *npj Digital Medicine* **6**, 131 (2023). https://doi.org:10.1038/s41746-023-00875-y

11 Shang, J., Ma, T., Xiao, C. & Sun, J. Pre-training of graph augmented transformers for medication recommendation. *arXiv preprint arXiv:1906.00346* (2019). https://doi.org:https://doi.org/10.48550/arXiv.1906.00346

12 Li, Y. *et al.* BEHRT: Transformer for Electronic Health Records. *Scientific Reports* **10**, 7155 (2020). https://doi.org:10.1038/s41598-020-62922-y

13 Rasmy, L., Xiang, Y., Xie, Z., Tao, C. & Zhi, D. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *npj Digital Medicine* **4**, 86 (2021). https://doi.org:10.1038/s41746-021-00455-y

14 Pang, C. *et al.* in *Proceedings of Machine Learning for Health* Vol. 158 (eds Roy Subhrajit *et al.*) 239--260 (PMLR, Proceedings of Machine Learning Research, 2021).

15 Li, Y. *et al.* Hi-BEHRT: Hierarchical Transformer-Based Model for Accurate Prediction of Clinical Events Using Multimodal Longitudinal Electronic Health Records. *IEEE Journal of Biomedical and Health Informatics* **27**, 1106-1117 (2023). https://doi.org:10.1109/JBHI.2022.3224727

16 Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018). https://doi.org:https://doi.org/10.48550/arXiv.1810.04805

17 Benson, T. *Principles of health interoperability HL7 and SNOMED.* (Springer Science & Business Media, 2012).

18 Liu, S., Wei, M., Moore, R., Ganesan, V. & Nelson, S. RxNorm: prescription for electronic drug information exchange. *IT Professional* **7**, 17-23 (2005). https://doi.org:10.1109/MITP.2005.122

19 McDonald, C. J. *et al.* LOINC, a Universal Standard for Identifying Laboratory Observations: A 5-Year Update. *Clinical Chemistry* **49**, 624-633 (2003). https://doi.org:10.1373/49.4.624

20 Vaswani, A. *et al.* Attention is all you need. *Advances in neural information processing systems* **30** (2017).

21 DeLong, E. R., DeLong, D. M. & Clarke-Pearson, D. L. Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics* **44**, 837-845 (1988). https://doi.org:10.2307/2531595

22 Wilson, E. B. Probable Inference, the Law of Succession, and Statistical Inference. *Journal of the American Statistical Association* **22**, 209-212 (1927). https://doi.org:10.1080/01621459.1927.10502953

23 Wolf, T. *et al.* Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771* (2019). https://doi.org:10.48550/arXiv.1910.03771

24 GOLDSTEIN, J. L. *et al.* Small bowel mucosal injury is reduced in healthy subjects treated with celecoxib compared with ibuprofen plus omeprazole, as assessed by video capsule endoscopy. *Alimentary Pharmacology & Therapeutics* **25**, 1211-1222 (2007). https://doi.org:https://doi.org/10.1111/j.1365-2036.2007.03312.x

25 Moore, A., Makinson, G. & Li, C. Patient-level pooled analysis of adjudicated gastrointestinal outcomes in celecoxib clinical trials: meta-analysis of 51,000 patients enrolled in 52 randomized trials. *Arthritis Research & Therapy* **15**, R6 (2013). https://doi.org:10.1186/ar4134

26 Cattaneo, M. Aspirin and Clopidogrel. *Arteriosclerosis, Thrombosis, and Vascular Biology* **24**, 1980-1987 (2004). https://doi.org:doi:10.1161/01.ATV.0000145980.39477.a9

27 Nalbandian, R. M., Mader, I. J., Barrett, J. L., Pearce, J. F. & Rupp, E. C. Petechiae, Ecchymoses, and Necrosis of Skin Induced by Coumarin Congeners: Rare, Occasionally Lethal Complication of Anticoagulant Therapy. *JAMA* **192**, 603-608 (1965). https://doi.org:10.1001/jama.1965.03080200021006

28 Stavorovsky, M., Lichtenstein, D. & Nissim, F. Skin Petechiae and Ecchymoses (Vasculitis) Due to Anticoagulant Therapy. *Dermatologica* **158**, 451-461 (2009). https://doi.org:10.1159/000250797

29 Brott, T. *et al.* Early Hemorrhage Growth in Patients With Intracerebral Hemorrhage. *Stroke*

28, 1-5 (1997). https://doi.org:doi:10.1161/01.STR.28.1.1

30 Melo, T. P., Pinto, A. N. & Ferro, J. M. Headache in intracerebral hematomas. *Neurology* **47**, 494-500 (1996). https://doi.org:doi:10.1212/WNL.47.2.494

31 Bozluolcay, M. *et al.* Hypercholesterolemia as one of the risk factors of intracerebral hemorrhage. *Acta Neurologica Belgica* **113**, 459-462 (2013). https://doi.org:10.1007/s13760-013-0222-6

32 Ma, C. *et al.* Low-density lipoprotein cholesterol and risk of intracerebral hemorrhage. *Neurology* **93**, e445-e457 (2019). https://doi.org:doi:10.1212/WNL.0000000000007853

33 Jackson, C. M., Esnouf, M. P. & Lindahl, T. L. A Critical Evaluation of the Prothrombin Time for Monitoring Oral Anticoagulant Therapy. *Pathophysiology of Haemostasis and Thrombosis* **33**, 43-51 (2003). https://doi.org:10.1159/000071641

34 Fujii, Y. *et al.* Hemostatic Activation in Spontaneous Intracerebral Hemorrhage. *Stroke* **32**, 883-890 (2001). https://doi.org:doi:10.1161/01.STR.32.4.883

35 Antovic, J., Bakic, M., Zivkovic, M., Ilic, A. & Blombäck, M. Blood coagulation and fibrinolysis in acute ischaemic and haemorrhagic (intracerebral and subarachnoid haemorrhage) stroke: does decreased plasmin inhibitor indicate increased fibrinolysis in subarachnoid haemorrhage compared to other types of stroke? *Scandinavian Journal of Clinical and Laboratory Investigation* **62**, 195-199 (2002). https://doi.org:10.1080/003655102317475452

36 Smith, A. *et al.* Which Hemostatic Markers Add to the Predictive Value of Conventional Risk Factors for Coronary Heart Disease and Ischemic Stroke? *Circulation* **112**, 3080-3087 (2005). https://doi.org:doi:10.1161/CIRCULATIONAHA.105.557132

37 Howie-Esquivel, J. & White, M. Biomarkers in Acute Cardiovascular Disease. *Journal of Cardiovascular Nursing* **23**, 124-131 (2008). https://doi.org:10.1097/01.Jcn.0000305072.49613.92

38 Aapro, M. *et al.* A randomized phase III study evaluating the efficacy and safety of NEPA, a fixed-dose combination of netupitant and palonosetron, for prevention of chemotherapy-induced nausea and vomiting following moderately emetogenic chemotherapy. *Annals of Oncology* **25**, 1328-1333 (2014). https://doi.org:https://doi.org/10.1093/annonc/mdu101

39 Gralla, R. J. *et al.* A phase III study evaluating the safety and efficacy of NEPA, a fixed-dose combination of netupitant and palonosetron, for prevention of chemotherapy-induced nausea and vomiting over repeated cycles of chemotherapy. *Annals of Oncology* **25**, 1333-1339 (2014). https://doi.org:https://doi.org/10.1093/annonc/mdu096

40 Carcinoembryonic Antigen. *Annals of Internal Medicine* **104**, 66-73 (1986). https://doi.org:10.7326/0003-4819-104-1-66 %m 3510056

41 Amri, R., Bordeianou, L. G., Sylla, P. & Berger, D. L. Impact of Screening Colonoscopy on Outcomes in Colon Cancer Surgery. *JAMA Surgery* **148**, 747-754 (2013). https://doi.org:10.1001/jamasurg.2013.8

42 Swartz, M. L. Drug Formus: NuLYTELY (PEG 3350, Sodium Chloride, Sodium Bicarbonate and Potassium Chloride for Oral Solution). *Gastroenterology Nursing* **14** (1992).

43 Witte, D. L., Kraemer, D. F., Dick, F. R. & Hamilton, H. Prediction of Bone Marrow Iron Findings from Tests Performed on Peripheral Blood. *American Journal of Clinical Pathology* **85**, 202-206 (1986). https://doi.org:10.1093/ajcp/85.2.202

44 Keng, M. K. & Sekeres, M. A. Febrile Neutropenia in Hematologic Malignancies. *Current Hematologic Malignancy Reports* **8**, 370-378 (2013). https://doi.org:10.1007/s11899-013-0171-4

45 ANDRÈS, E., KURTZ, J.-E. & MALOISEL, F. Nonchemotherapy drug-induced agranulocytosis: experience of the Strasbourg teaching hospital (1985–2000) and review of the literature. *Clinical & Laboratory Haematology* **24**, 99-106 (2002). https://doi.org:https://doi.org/10.1046/j.1365-2257.2002.00437.x

46 Wornow, M., Thapa, R., Steinberg, E., Fries, J. & Shah, N. Ehrshot: An ehr benchmark for few-shot evaluation of foundation models. *Advances in Neural Information Processing Systems* **36**, 67125-67137 (2023). https://doi.org:https://doi.org/10.48550/arXiv.2307.02028

47 Reinecke, I., Zoch, M., Reich, C., Sedlmayr, M. & Bathelt, F. The usage of OHDSI OMOP–a scoping review. *German Medical Data Sciences 2021: Digital Medicine: Recognize–Understand–Heal*, 95-103 (2021). https://doi.org:10.3233/SHTI210546

48 Lee, P. S. The Role of the Health Insurance Review and Assessment Service Referring to Changes in Health Care Environment. *HIRA* **1**, 81-84 (2021). https://doi.org:10.52937/hira.21.1.1.81

49 Lee, H., Ahn, H.-S., Kwon, S., Kang, H.-Y. & Han, E. Expert survey on real-world data utilization and real-world evidence generation for regulatory decision-making in drug lifecycle in Korea. *Clinical and Translational Science* **17**, e13801 (2024). https://doi.org:https://doi.org/10.1111/cts.13801

50 Seong, Y. *et al.* Incorporation of Korean Electronic Data Interchange Vocabulary into Observational Medical Outcomes Partnership Vocabulary. *hir* **27**, 29-38 (2021). https://doi.org:10.4258/hir.2021.27.1.29

51 Li, L., Fan, Y., Tse, M. & Lin, K.-Y. A review of applications in federated learning. *Computers & Industrial Engineering* **149**, 106854 (2020). https://doi.org:https://doi.org/10.1016/j.cie.2020.106854

52 Lee, D. Y. *et al.* Privacy-Preserving Federated Model Predicting Bipolar Transition in Patients With Depression: Prediction Model Development Study. *J Med Internet Res* **25**, e46165 (2023). https://doi.org:10.2196/46165

53 Lee, G. H. *et al.* Feasibility Study of Federated Learning on the Distributed Research Network of OMOP Common Data Model. *hir* **29**, 168-173 (2023). https://doi.org:10.4258/hir.2023.29.2.168

Author contributions

J.K., J.S.K., and K.K. conceptualized the study. J.K. conducted data curation, model development, and statistical analysis. K.K., M.G.K., and R.W.P. provided data and managed data security issues. J.H.L., T.K., C.C., and R.W.P. reviewed clinical validity of the study. J.K. wrote the original draft. J.S.K., J.H.L., and R.W.P. edited the draft.

Inclusion & Ethics

Competing interests

The authors declare no competing interests.

Materials & Correspondence

Kwangsoo Kim and Rae Woong Park.