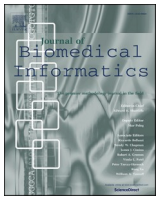
Journal of Biomedical Informatics 144 (2023) 104442

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin



Original Research

AD-BERT: Using pre-trained language model to predict the progression from mild cognitive impairment to Alzheimer's disease

Chengsheng Mao [a], Jie Xu [b,c], Luke Rasmussen [a], Yikuan Li [a], Prakash Adekkanattu [c], Jennifer Pacheco [a], Borna Bonakdarpour [d], Robert Vassar [d], Li Shen [e], Guoqian Jiang [f], Fei Wang [c], Jyotishman Pathak [c], Yuan Luo [a,*]

[a] Department of Preventive Medicine, Feinberg School of Medicine, Northwestern University, Chicago, IL, United States
[b] Department of Health Outcomes and Biomedical Informatics, University of Florida, Gainesville, FL, United States
[c] Weill Cornell Medicine, New York, NY, United States
[d] Department of Neurology, Feinberg School of Medicine, Northwestern University, Chicago, IL, United States
[e] Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, Philadelphia, PA, United States
[f] Mayo Clinic, Rochester, MN, United States

ARTICLE INFO

Keywords:
Alzheimer's disease
Mild cognitive impairment
Pre-trained language model
Electronic health records

ABSTRACT

Objective: We develop a deep learning framework based on the pre-trained Bidirectional Encoder Representations from Transformers (BERT) model using unstructured clinical notes from electronic health records (EHRs) to predict the risk of disease progression from Mild Cognitive Impairment (MCI) to Alzheimer's Disease (AD).
Methods: We identified 3657 patients diagnosed with MCI together with their progress notes from Northwestern Medicine Enterprise Data Warehouse (NMEDW) between 2000 and 2020. The progress notes no later than the first MCI diagnosis were used for the prediction. We first preprocessed the notes by deidentification, cleaning and splitting into sections, and then pre-trained a BERT model for AD (named AD-BERT) based on the publicly available Bio+Clinical BERT on the preprocessed notes. All sections of a patient were embedded into a vector representation by AD-BERT and then combined by global MaxPooling and a fully connected network to compute the probability of MCI-to-AD progression. For validation, we conducted a similar set of experiments on 2563 MCI patients identified at Weill Cornell Medicine (WCM) during the same timeframe.
Results: Compared with the 7 baseline models, the AD-BERT model achieved the best performance on both datasets, with Area Under receiver operating characteristic Curve (AUC) of 0.849 and F1 score of 0.440 on NMEDW dataset, and AUC of 0.883 and F1 score of 0.680 on WCM dataset.
Conclusion: The use of EHRs for AD-related research is promising, and AD-BERT shows superior predictive performance in modeling MCI-to-AD progression prediction. Our study demonstrates the utility of pre-trained language models and clinical notes in predicting MCI-to-AD progression, which could have important implications for improving early detection and intervention for AD.

1. Introduction

Alzheimer's Disease (AD) is the most prevalent progressive neurological disorder, and is the sixth leading cause of death in the United States since 2021 [1]. Approximately 5.8 million adults 65 years or older are living with AD currently, and by some estimates, that number will increase to 14 million adults by the year 2030 [2]. However, unfortunately, nearly 75% of patients suffering from AD remain undiagnosed globally due to stigma and lack of awareness [3].

The latest diagnostics guidelines classify AD into three stages based on patients' clinical symptoms [4]: preclinical AD [5], mild cognitive impairment (MCI) due to AD [6], and AD dementia [7]. People with MCI, which is marked by symptoms of abnormal memory and/or other thinking problems, may or may not progress to AD dementia, which is severe enough to impair a person's ability to function independently. Predicting the risk of MCI-to-AD progression is critical for clinical prognostication, risk stratification and early intervention [8]. Various approaches have been attempted to estimate the likelihood of disease

* Corresponding author at: 750 N. Lake Shore Drive, 11th floor, Chicago, IL 60611, United States.
E-mail address: yuan.luo@northwestern.edu (Y. Luo).

https://doi.org/10.1016/j.jbi.2023.104442
Received 24 March 2023; Received in revised form 13 June 2023; Accepted 7 July 2023
Available online 8 July 2023

progression from MCI to AD dementia using heterogeneous data modalities, including cerebrospinal fluid (CSF) biomarkers [9], magnetic resonance imaging (MRI) [10], positron emission tomography (PET) [11], genetics data [12,13] and the combination of them [14]. In addition, clinical measures such as Mini-Mental State Examination (MMSE) and Alzheimer's Disease Assessment Scale-Cognition (ADAS-Cog) have also been used for the prediction of MCI-to-AD progression [15–17]. Compared to the previous approaches, routinely collected clinical data from EHRs reflect real-world evidence. Specifically, clinical notes in EHRs contain rich information, such as family history, laboratory measurements, treatments, and self-report scores (e.g., MMSE), which may be analyzed to model the risk of disease progression from MCI-to-AD. Recently, studies have explored information from Electronic Health Records (EHRs) for AD-related research [18–20]. In particular, Fouladvand et al. [20] predicted the progression from cognitively unimpaired (CU) to MCI using demographics, clinical notes, and self-reported information. Wang et al. [19] proposed a deep learning model to detect the evidence of cognitive decline from clinical notes. However, research is still limited in analyzing EHRs to study MCI-to-AD progression prediction.

Recently, the Bidirectional Encoder Representations from Transformers (BERT) [21] model that was pretrained with contextualized embeddings has shown promising results in many NLP tasks, and has been extended to BioBERT [22] and Bio+Clinical BERT (BC-BERT) [23] for biomedical text and clinical narratives, respectively. These pre-trained language models have been used for the prediction or identification of unplanned readmission [24], heart failure [25], pancreatic cancer [25], acute kidney injury [26] and others. There were also studies using pre-trained language model for AD detection from spontaneous speech [27,28]. However, to the best of our knowledge, no prior study has applied a pre-trained language model to analyze clinical notes for Alzheimer's disease. Clinical notes related to MCI or AD usually have different terminologies in linguistic characteristics from clinical text for other diseases or general text. For example, the phrase "memory loss" appears more frequently in MCI or AD-related clinical notes than in general text. This motivates the need for specific BERT models for MCI-to-AD prediction.

Based on the above motivations, i.e., research is still limited in analyzing EHRs to study MCI-to-AD progression prediction and specific BERT models are needed to analyze EHR for MCI-to-AD prediction, we developed a deep learning framework based on BERT for MCI-to-AD risk prediction using clinical notes from EHRs, and validated it on an independent dataset from a different institute. In addition, the annual conversion rate from MCI to AD is typically reported as 10–15% [29,30], thus, the case group (i.e., the MCI patients who progressed to AD) usually has much fewer samples than the control group (i.e., the MCI patients who did not progress to AD), causing the class highly imbalanced. In a batch-based deep learning framework, the extremely imbalanced classes may cause a batch containing only control samples and no case samples, thus, making the training process fail. To address the class imbalance problem, we applied a stratified batch sampler to ensure that all batches have an equal ratio between case and control samples.

The main contributions of this article are summarized as follows. (1) We developed a deep learning framework called AD-BERT to predict the risk of MCI-to-AD progression using unstructured clinical notes from EHRs. We released the pre-trained model to hugging face https://huggingface.co/mocherson/AD-BERT/tree/main. (2) We applied a stratified batch sampler to address the class imbalance problem for batch training, ensuring that all batches have an equal ratio of case and control samples. (3) We validated AD-BERT on two real datasets and demonstrated its effectiveness for MCI-to-AD prediction, showing the utility of pre-trained language models and clinical notes in predicting MCI-to-AD progression, which could have important implications for improving early detection and intervention for AD.

2. Related work

Natural language processing (NLP) methods are broadly applied for the development of predictive models using clinical notes extracted from EHRs [31], some used the words directly as features and conventional machine learning methods to predict the patient outcomes [32–35], some enhanced conventional machine learning methods with Unified Medical Language System (UMLS) [36–42], some leveraged deep learning methods (e.g., convolutional neural networks) for disease onset and outcome prediction [43–50], some aimed to improve interoperability of NLP systems by anchoring on common data models [51,52]. In the context of AD, Fouladvand et al. [20] predicted the progression from cognitively unimpaired (CU) to MCI using demographics, clinical notes, and self-reported information and the best performing model (in this case, Long Short-Term Memory (LSTM)) achieved an AUC of 0.75 and F1 Score of 0.46. Wang et al. [19] proposed a deep learning model to detect the evidence of cognitive decline from clinical notes 4 years preceding the patient's first diagnosis of MCI. The proposed recurrent neural network demonstrated optimal predictive power of both AUC and AUPRC over 0.9. However, as far as we know, no prior study has attempted to predict the progression from MCI to AD using clinical narratives.

In recent years, transformer-based language models, e.g., BERT and Generative Pre-trained Transformer (GPT), have been pushing the state-of-the-art for heterogeneous NLP tasks including question answering, document classification, text generation and others. After observing the success of those models in the general domain, researchers released biomedical adapted variants of BERT, including BioBERT [22], ClinicalBERT [23], TCM-BERT [53], BlueBERT [54], by pre-training the attention-based transformer blocks with large scale clinical and biomedical corpora. These transformer-based models also achieved breakthrough results in phenotyping and clinical predictive tasks using EHR data. For instance, Venkatakrishnan et al. [55] investigated the associations between patients' pre-existing conditions and short-/long-term COVID-19 complications using a transformer-based model. The curated pre-existing conditions can be further used to predict the complications of COVID-19. In another study, Mao et al. [26] achieved better results in the prediction of AKI by pre-training a language model from clinical notes initialized by BERT weights. Rasmy et al. [25] proposed Med-BERT to adapt the BERT framework to the structured EHR domain with a clinical code as the basic token and substantially improved the prediction accuracy. Generative models that focus on generating creative and human-like outputs have also been investigated in the biomedical domain. Moradi et al. [56] investigated the ability of GPT [57], a recent generative model, for few-shot transfer learning in biomedical NLP tasks and showed that it could not perform as effectively as BioBERT. But the recent BioGPT [58] pre-trained on large-scale biomedical literature performed excellently in biomedical NLP tasks compared with its competitors. In the diagnosis of AD, transformer-based models have previously been utilized for detecting AD by distinguishing healthy individuals from those who are afflicted based on their speech patterns [27,28]. However, the application of transformer-based models to clinical notes for AD diagnosis has not been extensively explored yet.

3. Materials and methods

3.1. Study cohort

We identified a cohort of patients with MCI from the Northwestern Medicine Enterprise Data Warehouse (NMEDW) and Weill Cornell Medicine (WCM) using ICD-9 (331.83) and ICD-10 (G31.84) codes. Those who progressed to AD (identified by ICD-9 [331.0] and ICD-10 [G30.*]) were considered as the case group, and the control group was defined as MCI patients who have *not* yet been diagnosed with AD. We identified 396 cases and 3261 controls from NMEDW and 548 cases

and 2015 controls from WCM between 2000 and 2020. The exclusion criteria for NMEDW and WCM datasets are depicted in Fig. 1. The patients having only 1 encounter may indicate that their medical records are not complete in our healthcare system. It would be safer to exclude them than to count them as negative/control samples. All progress notes before the first encounter when a patient was diagnosed with MCI were collected for risk prediction.

3.2. Preprocessing

We preprocessed each progress note derived from the EHRs of all eligible patients as follows:

(1) Deidentification: Clinical notes contain legally Protected Health Information (PHI), such as patient names, addresses, and phone numbers, which should not be released to the public and should not be used for most research applications. We use the package Philter [59] to remove PHI for clinical notes.
(2) Cleaning: The cleaning step aims to preprocess the data and improve the model's ability to learn meaningful patterns from the text. Non-ASCII characters and extra spaces may not provide significant information for the model and could potentially introduce noise. They do not contribute to the semantic meaning of the text and can be safely removed without losing important information. Thus, following previous studies [26,33,60], we removed non-ASCII characters and extra spaces from the notes in the cleaning step.
(3) Splitting: Each note is split into sections by the newline character ('\n'). Each section is an input to the model independently for section representation. And the embedding vector for a patient generated by the global MaxPooling of all the section representations.

3.3. Framework

We first pretrained an AD-BERT model from BC-BERT [23] on the corpus of notes for all MCI patients from NMEDW dataset; the corpus contained about 37,000 clinical progress notes with an average length of 6250 characters. In the pretraining stage of AD-BERT, we followed the pretraining process of BERT with a loss function of the two unsupervised tasks, i.e., Masked Language Model (MLM) and Next Sentence Prediction (NSP), and pretrained AD-BERT from BC-BERT on the corpus of AD-related notes.

After AD-BERT was pretrained, we then fine-tuned AD-BERT for the MCI-to-AD prediction task using the training patients, and evaluated the model performance on an independent test set. The fine-tuning and testing were performed on both the NMEDW dataset and WCM dataset. Our framework is depicted in Fig. 2, where all the encounter notes of a patient are split into sections that are then used as an input to the pretrained AD-BERT. Different patients can have different numbers of notes and sections; AD-BERT generates a section embedding for each section independently; all the section embeddings of a patient are then combined to generate a vector representation for that individual by global MaxPooling. Then a fully connected linear layer appended with a sigmoid activation is employed to predict the probability of MCI-to-AD progression from the patient representation.

Global MaxPooling [61] aggregates the embeddings of a patient by selecting the maximum value for each feature. It provides a high-level summary of the patient notes by focusing on the most salient feature values across the entire sequence. It effectively captures the most dominant features and discards less important information. This pooling operation is commonly used as a dimensionality reduction technique, allowing subsequent layers to operate on a condensed representation of the input. By applying global MaxPooling to the section embeddings, the framework may be prioritizing the most relevant information from each section in a clinical note.

3.4. Configuration

We used the default configuration of BERT in *PyTorch transformers* [62] for AD-BERT pretraining[1], the default pretraining epoch is 3. For fine-tuning, each study cohort was randomly stratified into training and test sets by 8:2, and the training set was further split 1/5 as the validation set that was used to select the best model in the training process.

Deep learning models are usually trained based on batches so that only a batch of the training set is used to tune the parameters at a time by stochastic gradient descent [63]. Since the datasets were highly imbalanced, i.e., we had much more control samples than case samples, model training based on randomly selected batches would be biased to the control even if we employed a class-weighted loss function, because batches generated by random selection were likely to contain no case samples. Using a large batch size might reduce the bias, but a large batch would cause the out-of-memory issue for model training, especially for long notes. To address this issue, we applied a stratified batch sampler to ensure the same case/control ratio in all batches, and used a class-weighted loss function based on the case/control ratio. In our study, the batch size was set to 4 and the max sequence length was set to 32. We used a weighted cross-entropy loss with the class weights inversely proportional to the corresponding number in a batch. WCM data was used to validate our framework and the pretrained AD-BERT on an independent data source.

4. Results

We identified a total of 3657 MCI patients from NMEDW and 2563 patients from WCM. The summary statistics of the patients for NMEDW and WCM are listed in Table 1 and Table 2, respectively. NMEDW dataset contained 396 MCI patients (233 [58.8%] females; age mean [SD] 76.8 [9.0] years) who had progressed to AD. WCM dataset contained 548 MCI patients (311 [56.8%] females; age mean [SD] 74.4 [6.7] years) who had progressed to AD. The average conversion time is 723 days and 742 days in NMEDW and WCM, respectively. The conversion time is defined by the time length from the first MCI diagnosis date to the first AD diagnosis date (based on ICD-9/10 codes: 331.0, G30.*).

Besides the no-restrict prediction that is to predict whether an MCI patient will progress to AD without time restriction, we also considered the time windows of 6-month, 1-year, and 2-year to predict the MCI-to-AD progression, i.e., predicting whether an MCI patient would progress to AD in 6 months, 1 year, and 2 years, respectively. Samples in no-restrict prediction and *x*-month prediction are illustrated in Fig. 3. For no-restrict prediction Fig. 3(a), the case and control groups are distinct by the diagnosis condition, i.e., whether an AD diagnosis was found after MCI diagnosis. For a prediction in a certain time window of x months (x-month prediction) Fig. 3(b), besides the diagnosis condition (i.e., whether an AD diagnosis was found in x months after MCI diagnosis), we also considered the time length condition for the control group, i.e., the last encounter should be found after x months to ensure the patients have a conversion time of at least x months. We excluded the patients whose last encounter is in the time window and no AD diagnosis is detected. The patient counts for each setting are shown in Table 3.

We compared our model with several popular models for text classification, including BERT-based models such as basic BERT [21], BC-BERT [23], and BioBERT [22], bidirectional LSTM [64,65] with or without attention, CNN [60] and BOW with logistic regression [33]. We evaluated the predictive ability of a predictive model based on two metrics, F1 score and AUC. We conducted the prediction task on the independent test set using our framework as well as the baseline models. To ensure robustness, we resampled the prediction results 100 times

[1] https://github.com/huggingface/transformers/tree/v1.0.0/examples/lm_finetuning.

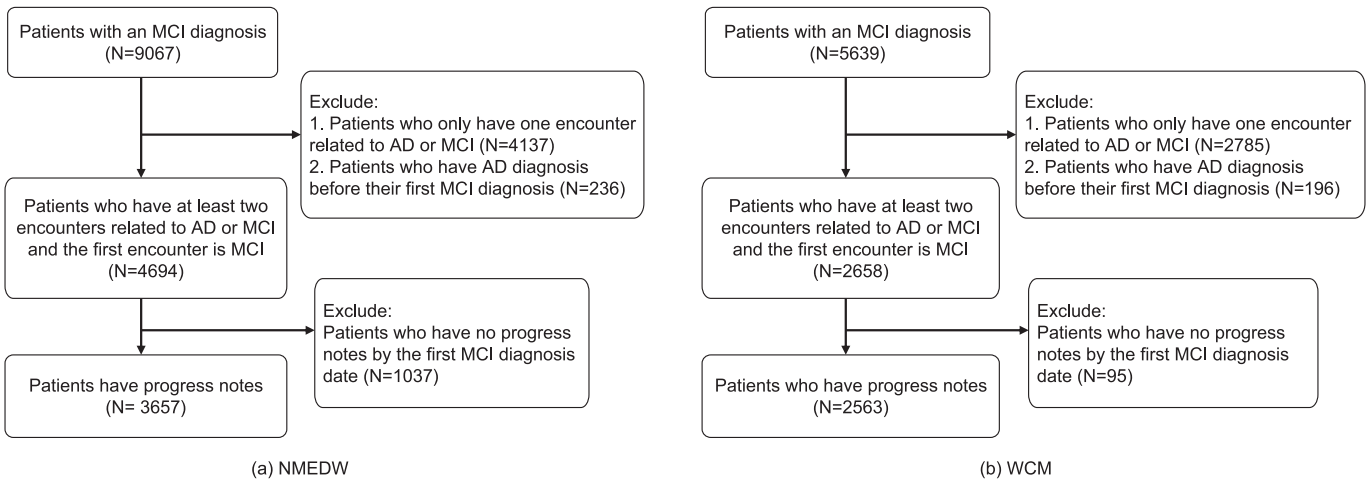


Fig. 1. Inclusion and exclusion criteria for the study cohorts for (a) NMEDW and (b) WCM.

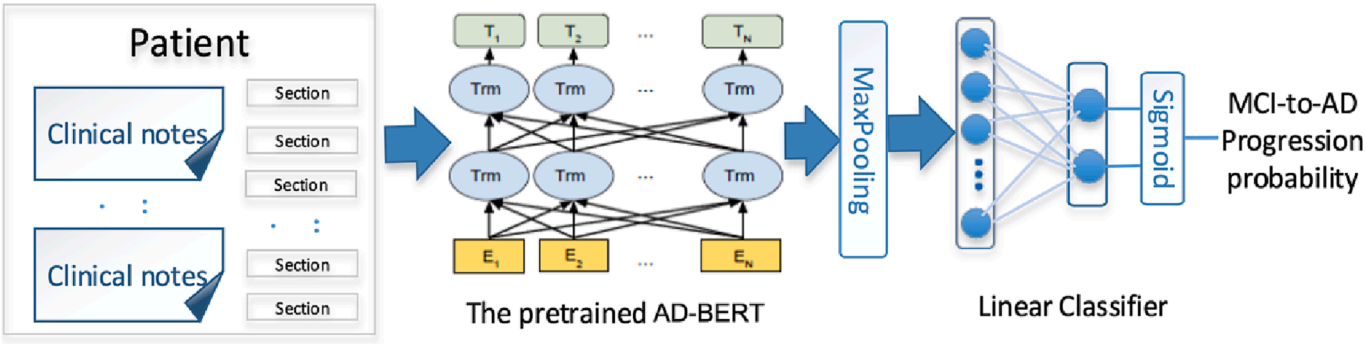


Fig. 2. The overview of our framework. The notes of a patient are split into sections, which are then fed to the pretrained AD-BERT model to generate a representation for each section. The patient representation is generated by global MaxPooling that aggregates all the section representations. Finally, a linear classifier combined with a sigmoid activation layer is used to predict probability of MCI-to-AD progression.

Table 1
Summary statistics of patients in NMEDW dataset.

Characteristic	Case (n = 396)	Control (n = 3261)
Age, Mean (SD), years	76.8 (9.0)	71.6 (15.8)
Sex, No. (%)		
Female	233 (58.8)	1680 (51.5)
Race, No. (%)		
White	311 (78.5)	2356 (72.2)
Black	31 (7.8)	401 (12.3)
Asian	3 (0.8)	79 (2.4)
Others & Unknown	51 (12.9)	425 (13.0)
Average conversion days	723 (MCI to AD)	521 (MCI lasts)

Table 2
Summary statistics of patients in WCM dataset.

Characteristic	Case (n = 548)	Control (n = 2015)
Age, Mean (SD), years	74.4 (6.7)	69.3 (13.0)
Sex, No. (%)		
Female	311 (56.8)	1089 (54.0)
Race, No. (%)		
White	325 (59.3)	1038 (51.5)
Black	28 (5.1)	154 (7.7)
Asian	9 (1.6)	55 (2.7)
Others & Unknown	186 (34.0)	768 (38.1)
Average conversion days	742 (MCI to AD)	511 (MCI lasts)

with replacement. For each sample, we calculated F1 score and AUC. The mean value of F1 and AUC are presented in Table 4. We also performed a one-tailed *t*-test between the best result and the second best result in Table 4. In most cases, AD-BERT achieved the highest performance and can significantly ($p = 0.05$) outperform the second best model. On NMEDW dataset, among all the models, AD-BERT achieved the best performance for all the settings, including no-restrict, 6-month, 1-year and 2-year predictions; the AUC for no-restrict prediction is 0.849 and the F1 score is 0.420, representing an increase of 4% in F1 and 1.6% in AUC compared with the best baseline mode (BC-BERT). The AD-BERT pretrained on NMEDW dataset also generalized well on the WCM dataset, achieving the best performance for all settings except 1-year prediction, with the no-restrict AUC 0.883 and the F1 score 0.680, representing an increase of 3.8% in F1 and 0.2% in AUC compared with the baseline model (BERT and BioBERT). Since F1 score depends on the share of actual positives, the discrepancy in F1 score between NMEDW and WCM is probably due to that we have less proportion of cases in NMEDW than in WCM.

We also find that deep learning models (e.g., BERT-based models, LSTM and CNN) achieve higher performance than BOW, indicating that only word counts in the notes cannot provide any information for MCI-to-AD prediction. AD-BERT pretrained on NMEDW can also perform well on WCM dataset, validating the generalizability of our framework.

We also investigated the prediction process to see what words the model will take more attention to make the prediction. Fig. 4 shows an example of attention visualization of AD-BERT. From Fig. 4, AD-BERT pays more attention to the terms like “memory”, “MCI” and “difficulty recalling dates” than others, which is reasonable as serious memory loss

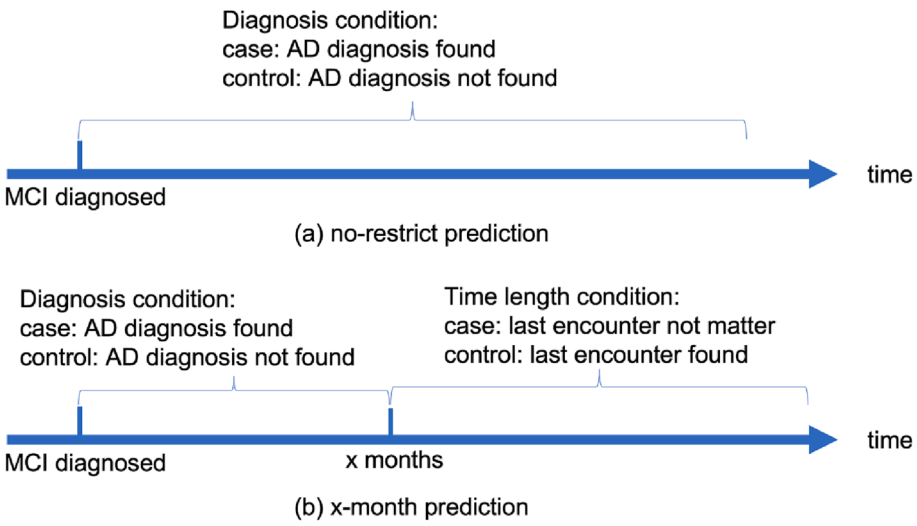


Fig. 3. The illustration of samples in case and control groups. (a) For no-restrict prediction, the case and control groups are differentiated by the AD diagnosis condition after MCI diagnosis, as reflected in all diagnostic records. (b) For x-month prediction, in addition to the AD diagnosis condition within x months after MCI diagnosis, we also enforce a time constraint on the control group by requiring the last encounter to occur after x months to ensure the patients have a conversion time of at least x months.

Table 3
Patient counts in each setting in NMEDW and WCM.

	NMEDW dataset			WCM dataset		
Conversion time	case	control	total	case	control	total
6 months	156	1295	1451	141	1463	1064
1 year	217	959	1176	221	1087	1308
2 years	277	615	892	338	644	982
no-restrict	396	3261	3657	548	2015	2563

is a typical symptom of AD.

5. Discussion

Recently, data from EHRs is being increasingly analyzed for AD-related research [17–19]. Our previous study has ever assessed longitudinal EHRs using machine learning (ML) methods to computationally derive probable subphenotypes for AD and related dementia [18], but research is limited in the application of pre-trained language model on clinical notes to study MCI-to-AD progression prediction.

Given that EHR is non-invasive and reflects real-world evidence compared to traditional biomarkers of AD, in this study, we developed a model to automatically predict the progression from MCI to AD using information and features derived from unstructured progress notes in EHR. Overall, our framework (AD-BERT) achieved the highest performance among commonly used NLP models. AD-BERT is a BERT model pretrained on AD-related clinical notes; it could effectively catch AD-related information in the clinical notes; thus AD-BERT is more effective than the models without pretraining on AD-related notes for AD-related tasks such as MCI-to-AD prediction.

In prior work, Young et al. [66] and Davatzikos et al. [14] achieved AUCs of 0.795 and 0.734, respectively, in the prediction of MCI-to-AD conversion using multiple biomarkers including MRI, PET, CSF and APOE. Zhang et al. [10] achieved an AUC of 0.888 in distinguishing MCI converter and non-converter using sMRI and fMRI data. In comparison to these studies, we note that (1) clinical notes used in our method are more routinely available for a larger sample of patients than biomarker data in previous studies; (2) these studies involved much fewer subjects (<300) than that in our study (3657 in NMEDW and 2563 in WCM); (3) these studies only considered the 3-year prediction, while we provided results for multiple conversion times including no-restriction prediction; (4) different from our study, baseline measures in these studies are not at the first time when MCI was diagnosed, which makes the exact conversion time not accurate. Nevertheless, our performance also validates the effectiveness of using clinical notes for MCI-to-AD prediction.

Furthermore, in this study, besides the no-restrict prediction, we also performed the MCI-to-AD prediction in 6 months, 1 year and 2 years. From the results, in terms of F1 score, we found that the performance of

Table 4
The performance of our model compared with other baselines for MCI to AD prediction. BiLSTM-att: bidirectional LSTM with attention. BOW+LR: bag-of-words and logistic regression. The best results are bolded. * denotes the best performance outperform the second best performance at a significance level of 0.05.

Dataset	Model	no-restrict		6-month		1-year		2-year	
		F1	AUC	F1	AUC	F1	AUC	F1	AUC
NMEDW	AD-BERT (ours)	**0.440***	**0.849***	**0.393***	**0.785***	**0.454***	**0.728***	**0.552***	**0.699***
	BC-BERT	0.420	0.835	0.362	0.771	0.422	0.711	0.537	0.633
	BioBERT	0.380	0.812	0.300	0.663	0.422	0.701	0.538	0.683
	BERT	0.379	0.784	0.301	0.729	0.419	0.707	0.543	0.682
	BiLSTM	0.332	0.727	0.283	0.659	0.353	0.600	0.474	0.642
	BiLSTM-att	0.360	0.754	0.300	0.718	0.358	0.605	0.432	0.619
	CNN	0.323	0.746	0.307	0.683	0.389	0.616	0.521	0.678
	BOW+LR	0.200	0.500	0.150	0.424	0.238	0.446	0.413	0.517
WCM	AD-BERT (ours)	**0.680***	**0.883**	**0.269**	**0.686***	0.372	0.666	**0.539***	**0.644***
	BC-BERT	0.495	0.743	0.260	0.674	**0.393**	**0.701**	0.464	0.488
	BioBERT	0.652	0.881	0.206	0.578	0.387	0.693	0.393	0.524
	BERT	0.655	0.877	0.253	0.673	0.380	0.671	0.423	0.581
	BiLSTM	0.455	0.682	0.165	0.492	0.247	0.498	0.459	0.548
	BiLSTM-att	0.487	0.750	0.209	0.569	0.332	0.594	0.517	0.595
	CNN	0.503	0.737	0.205	0.586	0.337	0.616	0.368	0.515
	BOW+LR	0.250	0.503	0.160	0.481	0.217	0.481	0.488	0.566

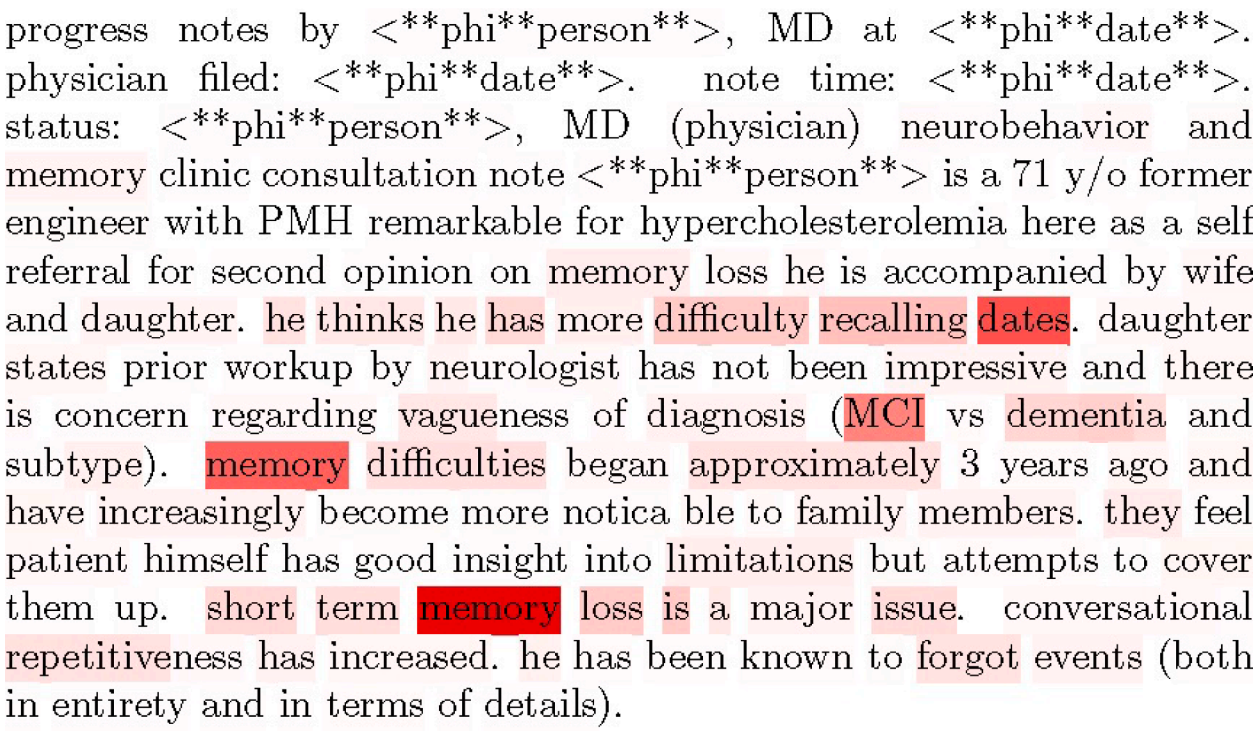


Fig. 4. Attention visualization of AD-BERT. The model pays more attention to the terms like "memory", "MCI" and "difficulty recalling dates" than others.

the 2-year prediction model was the best, and whereas, that of the 6-month model was the worst. This was expected because the prediction of a long-time window is easier than that of a short time window.

While NLP and ML technologies are very promising, a major challenge in broader adoption of these technologies is their portability across multiple EHR systems, i.e., developing methods that yield consistent results when applied to multiple, diverse settings. Coding practices, clinical practice patterns and physician behavior vary between clinical settings and will be reflected in clinical notes created at these sites. While previous studies have demonstrated varying levels of success in portability of NLP/ML technologies across institutions, recent work by Carrell et al. argues that there remain significant challenges in adapting NLP systems across multiple sites, which include assembling clinical corpora, managing diverse document structures and handling idiosyncratic linguistic expressions [67]. The fact that the AD-BERT model pretrained on NMEDW also performed well on WCM dataset further supports the generalizability of the framework.

Another innovation of this paper is the use of a stratified batch sampler to address the problem of data imbalance. Since the datasets in our study are quite imbalanced, the randomly selected batches usually contained no case samples, making the model biased toward the control class. In this paper, we applied a stratified batch sampler algorithm to ensure each batch contains an equal proportion of case samples. We also conducted the experiments using the randomly selected batches without stratified batch sampler, where all the BERT-based models output a 0 recall and NA precision for MCI-to-AD prediction, indicating that all the samples including case and control were predicted to the control group, making the model fail.

Like Bio-BERT and Clinical-BERT, AD-BERT shares the same architecture as the usual BERT but has different parameters. Specifically, AD-BERT is pre-trained on AD-related clinical notes, utilizing an initialization from BC-BERT. By leveraging the architecture of BERT and initializing with BC-BERT, AD-BERT likely inherits the powerful language modeling capabilities of BERT while focusing on the AD domain. This initialization helps AD-BERT start from a point of increased domain relevance, potentially leading to improved performance on AD-specific tasks compared to general-purpose BERT models.

Our model is different from the recent proposed Med-BERT in the input data modality. Med-BERT is designed specifically for structured EHR data where clinical codes are treated as tokens and patients are represented as sequences of clinical codes. On the other hand, our model focuses on unstructured clinical notes data, which comprises free-text narratives written by healthcare professionals, encompassing a wide range of information such as symptoms, diagnoses, treatments, and patient history. Med-BERT is well-suited for tasks that leverage structured EHR data and clinical codes. Conversely, our model is more suitable for tasks that require understanding and extracting information from unstructured clinical notes.

The study has several limitations. First, our model was pretrained on data from a single healthcare system (NMEDW) and validated using data from a single external dataset (WCM). Since clinical documentation and workflows vary significantly across healthcare systems, although our model pretrained on Northwestern data performed well on an external dataset (WCM), this does not guarantee the same level of performance on clinical notes from other healthcare systems. In our future work, additional validation studies will have to be conducted using clinical notes from multiple healthcare systems to pretrain a more generalized AD-BERT model.

Second, we only considered two stages (i.e., MCI and AD) in the AD progression process. AD progression is a complex long-term process, and our study to predict how likely an MCI patient will progress to AD represents a small fraction of AD progression. More effort will be made to find the risk factors that affect the AD progression in our future work, including normal to AD, normal to MCI, and progression between even more refined stages.

Third, the data usually has some records with missing values that could potentially affect the accuracy of the data. In our study, we excluded the patient encounters with missing ICD codes including those from other hospitals. This may exclude encounters with AD, resulting in assigning a case sample to the control group. This could be the reason that the case ratio in our no-restrict setting is lower than normal [29,30]. For the missing ICD codes, we can only improve the data quality in the database to reduce the missing rate. We also excluded the patients who have no progress notes when or before they were diagnosed with MCI.

This will reduce the overall sample size, not likely to affect the overall prediction results. Moreover, the patients with missing progress notes only accounted for a small fraction (~0.2 in NMEDW).

Fourth, we used ICD code to identify MCI and AD. ICD codes are primarily used for billing, and are not as reliable as the Clinical Dementia Rating (CDR) scale score or imaging biomarkers like amyloid and tau. However, CDR is only recorded in a limited subset of our patients and is not routinely contained in our dataset. Similarly, imaging biomarkers like amyloid and tau are also partially recorded in our datasets due to multiple reasons (e.g., patients may take imaging studies at other hospitals before coming to our healthcare systems). In this study, we want to prioritize the sensitivity of the ground truth. This approach will help us ensure that we do not miss any patients who may require further evaluation in downstream tasks. Therefore, we used ICD codes to identify MCI and AD similar to prior EHR-based studies [19,68–70].

Finally, we only considered the clinical notes for MCI-to-AD prediction in this study. Given that the biomarker data was also effective for AD prediction tasks, in our future work, we will try to combine clinical notes and biomarker data and other structured data to design an even more powerful multimodal approach for AD prediction.

6. Conclusion

Clinical notes contain rich information that may suggest disease progression. However, it is challenging to extract predictive information from unstructured notes. In this paper, we developed a deep learning framework based on BERT for MCI-to-AD risk prediction using clinical notes from EHRs, and validated it on an independent dataset from a different institute. In addition, we applied a stratified batch sampler to address the class imbalance issue between case and control. The deep learning framework using BERT models in this study may provide a solution for clinical note analysis for MCI-to-AD progression.

CRediT authorship contribution statement

Chengsheng Mao: Conceptualization, Methodology, Software, Data curation, Formal analysis, Investigation, Writing – original draft. **Jie Xu:** Data curation, Validation, Formal analysis, Investigation, Writing – original draft. **Luke Rasmussen:** Data curation, Formal analysis, Investigation, Writing – review & editing. **Yikuan Li:** Data curation, Formal analysis, Investigation, Writing – original draft. **Prakash Adekkanattu:** Validation, Writing – review & editing. **Jennifer Pacheco:** Validation, Writing – review & editing. **Borna Bonakdarpour:** Conceptualization, Writing – review & editing. **Robert Vassar:** Conceptualization, Writing – review & editing. **Li Shen:** Validation, Writing – review & editing. **Guoqian Jiang:** Validation, Writing – review & editing. **Fei Wang:** Validation, Supervision, Writing – review & editing. **Jyotishman Pathak:** Conceptualization, Resources, Supervision, Project administration, Funding acquisition, Writing – review & editing. **Yuan Luo:** Conceptualization, Resources, Supervision, Project administration, Funding acquisition, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The research is supported in part by US NIH grants R01GM105688 and R01MH121922.

References

[1] 2021 Alzheimer's disease facts and figures. Alzheimers Dement. 17(3) (Mar 2021) 327-406. doi:10.1002/alz.12328.
[2] W. Wong, Economic burden of Alzheimer disease and managed care considerations, Am. J. Manag. Care 26 (8 Suppl) (Aug 2020) S177–S183, https://doi.org/10.37765/ajmc.2020.88482.
[3] S. Gauthier, P. Rosa-Neto, J. Morais, C. Webster, World Alzheimer Report 2021: Journey through the diagnosis of dementia, Alzheimer's Disease International, 2021.
[4] C.R. Jack Jr., M.S. Albert, D.S. Knopman, et al., Introduction to the recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease, Alzheimers Dement. 7 (3) (May 2011) 257–262, https://doi.org/10.1016/j.jalz.2011.03.004.
[5] R.A. Sperling, P.S. Aisen, L.A. Beckett, et al., Toward defining the preclinical stages of Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease, Alzheimers Dement. 7 (3) (2011) 280–292.
[6] M.S. Albert, S.T. DeKosky, D. Dickson, et al., The diagnosis of mild cognitive impairment due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease, Alzheimers Dement. 7 (3) (2011) 270–279.
[7] G.M. McKhann, D.S. Knopman, H. Chertkow, et al., The diagnosis of dementia due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease, Alzheimers Dement. 7 (3) (2011) 263–269.
[8] M.S. Chong, S. Sahadevan, Preclinical Alzheimer's disease: diagnosis and prediction of progression, Lancet Neurol. 4 (9) (2005) 576–579.
[9] O. Hansson, S. Lehmann, M. Otto, H. Zetterberg, P. Lewczuk, Advantages and disadvantages of the use of the CSF Amyloid β (Aβ) 42/40 ratio in the diagnosis of Alzheimer's Disease, Alzheimers Res. Ther. 11 (1) (2019) 1–15.
[10] T. Zhang, Q. Liao, D. Zhang, et al., Predicting MCI to AD Conversation Using Integrated sMRI and rs-fMRI: Machine Learning and Graph Theory Approach, Front. Aging Neurosci. 13 (2021), 688926.
[11] L. Teng, Y. Li, Y. Zhao, et al., Predicting MCI progression with FDG-PET and cognitive scores: a longitudinal study, BMC Neurol. 20 (1) (2020) 148, https://doi.org/10.1186/s12883-020-01728-x.
[12] A.J. Saykin, L. Shen, X. Yao, et al., Genetic studies of quantitative MCI and AD phenotypes in ADNI, Progress, Opportunities, Plans 11 (7) (2015) 792–814.
[13] D. Shigemizu, S. Akiyama, S. Higaki, et al., Prognosis prediction model for conversion from mild cognitive impairment to Alzheimer's disease created by integrative analysis of multi-omics data, Alzheimers Res. Ther. 12 (1) (2020) 145, https://doi.org/10.1186/s13195-020-00716-0.
[14] C. Davatzikos, P. Bhatt, L.M. Shaw, K.N. Batmanghelich, J.Q. Trojanowski, Prediction of MCI to AD conversion, via MRI, CSF biomarkers, and pattern classification, Neurobiol. Aging 32 (12) (2011) e19-2322. e27.
[15] I.O. Korolev, L.L. Symonds, A.C. Bozoki, Predicting Progression from Mild Cognitive Impairment to Alzheimer's Dementia Using Clinical, MRI, and Plasma Biomarkers via Probabilistic Pattern Classification, PLoS One 11 (2) (2016) e0138866.
[16] R.O. Roberts, D.S. Knopman, M.M. Mielke, et al., Higher risk of progression to dementia in mild cognitive impairment cases who revert to normal, Neurology 82 (4) (2014) 317–325, https://doi.org/10.1212/WNL.0000000000000055.
[17] K. Tifratene, P. Robert, A. Metelkina, C. Pradier, J.F. Dartigues, Progression of mild cognitive impairment to dementia due to AD in clinical settings, Neurology 85 (4) (2015) 331–338.
[18] J. Xu, F. Wang, Z. Xu, et al., Data-driven discovery of probable Alzheimer's disease and related dementia subphenotypes using electronic health records, Learning Health Systems 4 (4) (2020), e10246.
[19] L. Wang, J. Laurentiev, J. Yang, et al., Development and Validation of a Deep Learning Model for Earlier Detection of Cognitive Decline From Clinical Notes in Electronic Health Records, JAMA Netw. Open 4 (11) (2021) e2135174.
[20] S. Fouladvand, M.M. Mielke, M. Vassilaki, J.S. Sauver, R.C. Petersen, S. Sohn, Deep Learning Prediction of Mild Cognitive Impairment using Electronic Health Records, IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (2019) 799–806.
[21] J.-D.-M.-W.-C. Kenton, L.K.B.E.R.T. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, Proceedings of NAACL-HLT (2019) 4171–4186.
[22] J. Lee, W. Yoon, S. Kim, et al., BioBERT: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics 36 (4) (2020) 1234–1240.
[23] E. Alsentzer, J.R. Murphy, W. Boag, et al., Publicly available clinical BERT embeddings, (2019) arXiv preprint arXiv:1904033.
[24] K. Huang, J. Altosaar, R. Ranganath, Clinicalbert: Modeling clinical notes and predicting hospital readmission, (2019) arXiv preprint arXiv:190405342.
[25] L. Rasmy, Y. Xiang, Z. Xie, C. Tao, D. Zhi, Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction, NPJ. Digit. Med. 4 (1) (2021) 86, https://doi.org/10.1038/s41746-021-00455-y.
[26] C. Mao, L. Yao, Y. Luo, A Pre-trained Clinical Language Model for Acute Kidney Injury, IEEE International Conference on Healthcare Informatics (ICHI) (2020) 1–2.
[27] J.H. Yuan, Y.C. Bian, X.Y. Cai, J.J. Huang, Z. Ye, K. Church, Disfluencies and Fine-Tuning Pre-trained Language Models for Detection of Alzheimer's Disease,

Interspeech. 2162–2166 (2020), https://doi.org/10.21437/Interspeech.2020-2516.
[28] F. Agbavor, H. Liang, Predicting dementia from spontaneous speech using large language models, PLOS Digit Health 1 (12) (Dec 2022) e0000168.
[29] R.C. Petersen, G.E. Smith, S.C. Waring, R.J. Ivnik, E.G. Tangalos, E. Kokmen, Mild cognitive impairment: clinical characterization and outcome, Arch. Neurol. 56 (3) (Mar 1999) 303–308, https://doi.org/10.1001/archneur.56.3.303.
[30] A.J. Mitchell, M. Shiri-Feshki, Temporal trends in the long term risk of progression of mild cognitive impairment: a pooled analysis, J. Neurol. Neurosurg. Psychiatry 79 (12) (Dec 2008) 1386–1391, https://doi.org/10.1136/jnnp.2007.142679.
[31] Z. Zeng, Y. Deng, X. Li, T. Naumann, Y. Luo, Natural Language Processing for EHR-Based Computational Phenotyping, IEEE/ACM Trans. Comput. Biol. Bioinf. 16 (1) (2018) 139–153, https://doi.org/10.1109/TCBB.2018.2849968.
[32] B.J. Marafino, M. Park, J.M. Davies, et al., Validation of prediction models for critical care outcomes using natural language processing of electronic health record data, JAMA Netw. Open 1 (8) (2018) e185097–e.
[33] Y. Li, L. Yao, C. Mao, A. Srivastava, X. Jiang, Y. Luo, Early Prediction of Acute Kidney Injury in Critical Care Setting Using Clinical Notes, IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (2018) 683–686.
[34] S. Pakhomov, N. Shah, P. Hanson, S. Balasubramaniam, S.A. Smith, Automatic quality of life prediction using electronic medical records, Am. Med. Informatics Assoc. (2008) 545.
[35] J. Shin, Y. Li, Y. Luo, Early Prediction of Mortality in Critical Care Setting in Sepsis Patients Using Structured Features and Unstructured Clinical Notes, IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (2021) 2885–2890.
[36] D. Van Le, J. Montgomery, K.C. Kirkby, J. Scanlan, Risk prediction using natural language processing of electronic mental health records in an inpatient forensic psychiatry setting, J. Biomed. Inform. 86 (2018) 49–58.
[37] Z. Zeng, L. Yao, A. Roy, et al., Identifying Breast Cancer Distant Recurrences from Electronic Health Records Using Machine Learning, J. Healthcare Inform. Res. (2019) 1–17, https://doi.org/10.1007/s41666-019-00046-3.
[38] S. Sabra, K.M. Malik, M. Alobaidi, Prediction of venous thromboembolism using semantic and sentiment analyses of clinical narratives, Comput. Biol. Med. 94 (2018) 1–10.
[39] Z. Zeng, S. Espino, A. Roy, et al., Using natural language processing and machine learning to identify breast cancer local recurrence, BMC Bioinf. 19 (17) (2018) 65–74.
[40] J. Ye, L. Yao, J. Shen, R. Janarthanam, Y. Luo, Predicting mortality in critically ill patients with diabetes using machine learning and clinical notes, BMC Med. Inf. Decis. Making 20 (11) (2020) 1–7.
[41] L.-w. Lehman, M. Saeed, W. Long, J. Lee, R. Mark, Risk stratification of ICU patients using topic models inferred from unstructured progress notes, American Medical Informatics Association, 2012, p. 505.
[42] Z. Zeng, X. Li, S. Espino, et al., Contralateral Breast Cancer Event Detection Using Natural Language Processing, American Medical Informatics Association, 2017, pp. 1885–1892.
[43] N.W. Sterling, R.E. Patzer, M. Di, J.D. Schrager, Prediction of emergency department patient disposition based on natural language processing of triage notes, Int. J. Med. Inf. 129 (2019) 184–188.
[44] H. Wang, Y. Li, S.A. Khan, Y. Luo, Prediction of breast cancer distant recurrence using natural language processing and knowledge-guided convolutional neural network, Artificial intelligence in medicine (2020) 101977.
[45] B. Tahayori, N. Chini-Foroush, H. Akhlaghi, Advanced natural language processing technique to predict patient disposition based on emergency triage notes, Emerg. Med. Australas. 33 (3) (2021) 480–484.
[46] L. Yao, C. Mao, Y. Luo, Clinical text classification with rule-based features and knowledge-guided convolutional neural networks, BMC Med. Inf. Decis. Making 19 (Suppl 3) (2019) 71.
[47] É. Arnaud, M. Elbattah, M. Gignon, G. Dequen, Deep learning to predict hospitalization at triage: Integration of structured data and unstructured text, IEEE International Conference on Big Data (Big Data) (2020) 4836–4841.
[48] M. Sun, J. Baron, A. Dighe, et al., Early Prediction of Acute Kidney Injury in Critical Care Setting Using Clinical Notes and Structured Multivariate Physiological Measurements, Stud. Health Technol. Inform. 264 (2019) 368–372.
[49] C.-H. Chen, J.-G. Hsieh, S.-L. Cheng, Y.-L. Lin, P.-H. Lin, J.-H. Jeng, Emergency department disposition prediction using a deep neural network with integrated clinical narratives and structured data, Int. J. Med. Inf. 139 (2020), 104146.
[50] Y. Li, H. Wang, Y. Luo, A Comparison of Pre-trained Vision-and-Language Models for Multimodal Representation Learning across Medical Images and Reports, IEEE international conference on bioinformatics and biomedicine (BIBM) (2020) 1999–2004.
[51] J.R. Brown, I.M. Ricket, R.M. Reeves, et al., Information Extraction From Electronic Health Records to Predict Readmission Following Acute Myocardial Infarction: Does Natural Language Processing Using Clinical Notes Improve Prediction of Readmission? J. Am. Heart Assoc. 11 (7) (2022) e024198.
[52] H. Sharma, C. Mao, Y. Zhang, et al., Developing a portable natural language processing based phenotyping system, BMC Med. Inf. Decis. Making 19 (Suppl 3) (2019) 78.
[53] L. Yao, Z. Jin, C. Mao, Y. Zhang, Y. Luo, Traditional Chinese medicine clinical records classification with BERT and domain specific corpora, Journal of the American Medical Informatics Association 26 (12) (2019) 1632–1636.
[54] Y. Peng, S. Yan, Z. Lu, Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets, Proceedings of the 18th BioNLP Workshop and Shared Task (2019) 58–65.
[55] A.J. Venkatakrishnan, C. Pawlowski, D. Zemmour, et al., Mapping each pre-existing condition's association to short-term and long-term COVID-19 complications, npj digital medicine 4 (1) (2021) 117.
[56] M. Moradi, K. Blagec, F. Haberl, M. Samwald, Gpt-3 models are poor few-shot learners in the biomedical domain, (2021) arXiv preprint arXiv:210902555.
[57] T. Brown, B. Mann, N. Ryder, et al., Language models are few-shot learners, Adv. Neural Inf. Proces. Syst. 33 (2020) 1877–1901.
[58] R. Luo, L. Sun, Y. Xia, et al., BioGPT: generative pre-trained transformer for biomedical text generation and mining, Brief. Bioinform. 23 (6) (2022).
[59] B. Norgeot, K. Muenzen, T.A. Peterson, et al., Protected Health Information filter (Philter): accurately and securely de-identifying free-text clinical notes, NPJ Digit Med. 3 (2020) 57, https://doi.org/10.1038/s41746-020-0258-y.
[60] Y. Kim, Convolutional Neural Networks for Sentence Classification, Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) (2014).
[61] R. Nirthika, S. Manivannan, A. Ramanan, R.X. Wang, Pooling in convolutional neural networks for medical image analysis: a survey and an empirical study, Neural Comput. Appl. 34 (7) (Apr 2022) 5321–5347, https://doi.org/10.1007/s00521-022-06953-8.
[62] T. Wolf, L. Debut, V. Sanh, et al., Transformers: State-of-the-art natural language processing, Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations (2020) 38–45.
[63] L. Bottou, Stochastic gradient descent tricks. Neural networks: Tricks of the trade, Springer, 2012, pp. 421–436.
[64] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780, https://doi.org/10.1162/neco.1997.9.8.1735.
[65] P. Liu, X. Qiu, X. Huang, Recurrent neural network for text classification with multi-task learning, Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI) (2016) 2873–2879.
[66] J. Young, M. Modat, M.J. Cardoso, et al., Accurate multimodal probabilistic prediction of conversion to Alzheimer's disease in patients with mild cognitive impairment, NeuroImage: Clin. 2 (2013) 735–745.
[67] D.S. Carrell, R.E. Schoen, D.A. Leffler, et al., Challenges in adapting existing clinical natural language processing systems to multiple, diverse health care settings, J. Am. Med. Inform. Assoc. 24 (5) (Sep 2017) 986–991, https://doi.org/10.1093/jamia/ocx039.
[68] M.F. Iulita, D. Garzon Chavez, M. Klitgaard Christensen, et al., Association of Alzheimer Disease With Life Expectancy in People With Down Syndrome, JAMA Netw. Open 5 (5) (May 2 2022) e2212910.
[69] J.S. Fang, P.Y. Zhang, Y.D. Zhou, et al., Endophenotype-based in silico network medicine discovery combined with insurance record data mining identifies sildenafil as a candidate drug for Alzheimer's disease, Nature Aging 1 (12) (2021), https://doi.org/10.1038/s43587-021-00138-z, 1175-1188.
[70] J.L. Xu, C.S. Mao, Y. Hou, et al., Interpretable deep learning translation of GWAS and multi-omics findings to identify pathobiology and drug repurposing in Alzheimer's disease, Cell Rep. 41(9) (2022), doi:ARTN 111717 10.1016/j.celrep.2022.111717.