Journal of the American Medical Informatics Association, 30(2), 2023, 340–347
https://doi.org/10.1093/jamia/ocac225
Advance Access Publication Date: 30 November 2022
Research and Applications

Research and Applications

A comparative study of pretrained language models for long clinical text

Yikuan Li [1], Ramsey M. Wehbe [2,3], Faraz S. Ahmad [1,2,3], Hanyin Wang [1], and Yuan Luo [1]

[1]Division of Health and Biomedical Informatics, Department of Preventive Medicine, Feinberg School of Medicine, Northwestern University, Chicago, Illinois, USA, [2]Division of Cardiology, Department of Medicine, Feinberg School of Medicine, Northwestern University, Chicago, Illinois, USA, [3]Bluhm Cardiovascular Institute Center for Artificial Intelligence, Northwestern Medicine, Chicago, Illinois, USA

Corresponding Author: Yuan Luo, Division of Health and Biomedical Informatics, Department of Preventive Medicine, Feinberg School of Medicine, Northwestern University, Chicago, IL 60611, USA; yuan.luo@northwestern.edu

Received 10 August 2022; Revised 6 November 2022; Editorial Decision 8 November 2022; Accepted 14 November 2022

ABSTRACT

Objective: Clinical knowledge-enriched transformer models (eg, ClinicalBERT) have state-of-the-art results on clinical natural language processing (NLP) tasks. One of the core limitations of these transformer models is the substantial memory consumption due to their full self-attention mechanism, which leads to the performance degradation in long clinical texts. To overcome this, we propose to leverage long-sequence transformer models (eg, Longformer and BigBird), which extend the maximum input sequence length from 512 to 4096, to enhance the ability to model long-term dependencies in long clinical texts.

Materials and methods: Inspired by the success of long-sequence transformer models and the fact that clinical notes are mostly long, we introduce 2 domain-enriched language models, Clinical-Longformer and Clinical-BigBird, which are pretrained on a large-scale clinical corpus. We evaluate both language models using 10 baseline tasks including named entity recognition, question answering, natural language inference, and document classification tasks.

Results: The results demonstrate that Clinical-Longformer and Clinical-BigBird consistently and significantly outperform ClinicalBERT and other short-sequence transformers in all 10 downstream tasks and achieve new state-of-the-art results.

Discussion: Our pretrained language models provide the bedrock for clinical NLP using long texts. We have made our source code available at https://github.com/luoyuanlab/Clinical-Longformer, and the pretrained models available for public download at: https://huggingface.co/yikuan8/Clinical-Longformer.

Conclusion: This study demonstrates that clinical knowledge-enriched long-sequence transformers are able to learn long-term dependencies in long clinical text. Our methods can also inspire the development of other domain-enriched long-sequence transformers.

Key words: clinical natural language processing, text classification, named entity recognition, question answering, natural language inference

INTRODUCTION

Transformer-based models have been wildly successful in setting state-of-the-art benchmarks on a broad range of natural language processing (NLP) tasks, including question answering (QA), document classification, machine translation, text summarization, and others.[1–3] These successes have been replicated in the clinical and biomedical domain via pretraining language models using large-scale clinical or biomedical corpora, then fine-tuning on a variety of clinical or biomedical downstream tasks, including computational phenotyping,[4] automatical ICD (International Classification of Diseases) coding,[5] knowledge graph completion,[6] and clinical QA.[7]

The self-attention mechanism[8] is one of the most critical components that lead to the success of transformer-based models, which allows each token in the input sequence to independently interact with every other token in the sequence in parallel. However, the memory consumption of the self-attention mechanism grows quadratically with sequence length, resulting in impracticable training time, and easily reaching the memory limits of modern GPUs (graphic processing units). Consequently, transformer-based models that leverage a complete self-attention mechanism, such as BERT and RoBERTa, typically have an input sequence length limit of 512 tokens. To deal with this limit when modeling long texts using transformer-based models, the input sequence shall be either truncated to the first 512 tokens or processed via a sliding window of 512 tokens with or without overlap. If the latter method is applied to a document-level classification task, an aggregation operation will be added to yield the final output from multiple snippets. Both methods ignore long-term dependencies spanning over 512 tokens and may achieve suboptimal results due to information loss. Additionally, this input token limitation of the self-attention mechanism could impact language model pretraining and then be amplified to downstream tasks. In clinical NLP, transformer-based modeling approaches have also encountered this limitation.[9] For example, the discharge summaries in MIMIC-III, which are often used to predict clinically meaningful events like hospital readmission[10] or mortality,[11] have 2984 tokens (1435 words) on average, far exceeding the 512 token limits of most full attention-based transformer models.

Recently, investigators have developed novel variants of transformers specifically for long sequences that reduce memory usage from quadratic to linear scale of the sequence length.[12–14] The core idea behind these models is to replace the full attention mechanism with a sparse attention mechanism, which is typically a blend of sliding windows and reduced global attention. These models are capable of processing up to 4096 tokens and have empirically boosted performance on NLP tasks, including QA as well as text summarization. However, to the best of our knowledge, long-sequence transformers in the clinical and biomedical domains have not yet been systematically explored. The purpose of this article is to examine the adaptability of these long-sequence models to a series of clinical NLP tasks. We make the following contributions:

- We leverage large-scale clinical notes to pretrain 2 new language models, namely Clinical-Longformer and Clinical-BigBird.
- We demonstrate that both Clinical-Longformer and Clinical-BigBird improve the performance of a variety of downstream clinical NLP datasets, including QA, named entity recognition, and document classification tasks.

BACKGROUND AND SIGNIFICANCE

Clinical and biomedical transformers

Transformer-based models, especially BERT,[2] can be enriched with clinical and biomedical knowledge through pretraining on large-scale clinical and biomedical corpora. These domain-enriched models, for example, BioBERT[15] pretrained on biomedical publications and ClinicalBERT[16] pretrained on clinical narratives, set state-of-the-art benchmarks on downstream clinical and biomedical NLP tasks. Inspired by the success of these domain-enriched models, more pretrained models were released to boost the performance of NLP models when applied to specific clinical scenarios. For example, Smit et al[17] proposed CheXbert to annotate thoracic disease findings from radiology reports, which outperformed previous rule-based labelers with statistical significance. The model was pretrained using a combination of human-annotated and machine-annotated radiology reports. He et al[18] introduced DiseaseBERT, which infused disease knowledge to the BERT model by pretraining on a series of disease description passages that were constructed from Wikipedia and MeSH terms. DiseaseBERT achieved superior results on consumer health QA tasks compared with BERT and ClinicalBERT. Michalopoulos et al[19] proposed UmlsBERT, which used the Unified Medical Language System Metathesaurus to augment the domain knowledge learning ability of ClinicalBERT. Zhou et al[20] developed CancerBERT to extract breast cancer-related concepts from clinical notes and pathology reports. Agrawal et al[21] leveraged order contrastive pretraining on longitudinal data to tackle the difficulty when only a small proportion of the clinical notes were annotated. However, all models mentioned above were built on the vanilla BERT architecture, which has a limitation of 512 tokens in the input sequence length. This limitation may result in the information loss of long-term dependencies in the training processes.

Transformers for long sequences

Various attention mechanisms have been proposed to handle the large memory consumption of the attention operations in the vanilla transformer architecture. Transformer-XL[22] segmented a long sequence into multiple small chunks and then learned long-term dependencies with a left-to-right segment-level recurrence mechanism. Transformer-XL learns 5.5 times longer dependencies than the vanilla transformer models but loses the advantage of bidirectional representation of BERT-like models. In another study, Reformer[23] applied 2 techniques to reduce the complexity of transformer architecture by replacing dot-product attention operation with locality-sensitive hashing and sharing the activation function among layers. Reformer was able to process longer sequences at a faster speed and be more memory efficient. However, this enhancement improves space, time, and memory efficiency, but not accuracy on specific tasks. Almost simultaneously, Longformer[13] and BigBird[14] were proposed to drastically alleviate the memory consumption of transformer models by replacing the pairwise full attention mechanisms with a combination of sliding window attention and global attention mechanisms. They are slightly different regarding the implementation and configuration of the global and local attention mechanism, where BigBird introduces additional contrastive predictive coding to train global tokens.[14] Both models support input sequences up to 4.096 tokens long (8 times the input sequence limit of BERT) and significantly improve performance on long-text QA and summarization tasks. However, the adaptability of these

long-sequence transformers to the clinical and biomedical fields, where document length mostly exceeds the limits of BERT-like models, has not been investigated.

MATERIALS AND METHODS

In this section, we first introduce the clinical dataset we use as the pretraining corpus, followed by the pretraining processes for Clinical-Longformer and Clinical-BigBird. Next, we enumerate the downstream tasks we use to compare our long-sequence models with the short-sequence models. We also provide the technical details of pretraining and fine-tuning for the purposes of reproducing our results. The entire pipeline can be found in Figure 1.

Datasets

Similar to Huang et al[10] and Alsentzer et al,[16] we leverage approximately 2 million clinical notes extracted from the MIMIC-III[24] dataset, which is the most extensive publicly available electronic health records (EHR) dataset that contains clinical narratives of over 40 000 patients admitted to the intensive care units (ICUs). We only apply minimal preprocessing steps, including (1) to remove all the deidentification placeholders from the clinical notes that were generated to protect the protected health information (PHI); (2) to remove all characters other than alphanumerics and punctuation marks; (3) to convert all alphabetical characters to lower cases, and (4) to strip extra white spaces. We believe that complicated preprocessing in the pretraining stage may not improve downstream performance but will sacrifice the generalizability of language models and significantly increase training time.

Pretraining

Longformer[13] and BigBird[14] are the 2 best-performing transformer models that are designed for long input sequences. Both models extend the maximum input sequence length to 4096 tokens, which is 8× the limit of conventional transformer-based models, by introducing localized sliding windows and global attention mechanisms to reduce the computational expenses of full self-attention mechanisms. The differences between the 2 models are how the global attention is realized and the selection of loss function in fine-tuning.[13] BigBird also contains some random localized attention operations. The reported performance difference between the 2 models on downstream tasks is minimal.[14] Therefore, we seek to pretrain both models and compare their performance on clinical NLP tasks. We refer readers to the original papers of Longformer[13] and BigBird[14] for more technical details.

We initialize Clinical-Longformer and Clinical-BigBird from the pretrained weights of the base version of Longformer and the internal transformer construction (ITC) version of BigBird, respectively. Although the extended transformer construction (ETC) version of BigBird may have superior performance, HuggingFace (the largest community for sharing open-source pretrained transformer models) only provides the implementation and the pretrained checkpoints of the ITC version. The difference between ITC and ETC versions is that in ITC version some existing tokens are made "global" and attend over the entire sequence, while the ETC version introduces additionally "global" tokens such as CLS. Byte-level Byte-Pair-Encoding[25] is applied to tokenize the clinical corpus. Both models are distributed in parallel to six 32GB Tesla V100 GPUs. FP16 precision is enabled to accelerate training. Batch size is 18 for Clinical-Longformer and 12 for Clinical-Bigbird, which are the upper limits under 6 32GB GPUs. We pretrain Clinical-Longformer for 200 000 steps and Clinical-BigBird for 300 000 steps, which ensures that each clinical note is seen equal times by the 2 models. The learning rates are 3e−5 for both models, the same as the learning rate used in the pretraining of Longformer. The entire pretraining process takes more than 2 weeks for each model.

To evaluate the performance of pretraining, we create a testing set that contains 1000 documents that are also from MIMIC-III but have not been used as the pretraining corpora. Each document in the testing set is truncated to 512 tokens long. We randomly select 10% tokens from each document and replace them with a mask token. We compare our 2 pretrained models with the short-sequence models in filling in the masked tokens using context. We report the perplexity score and top 5 accuracies in filling in the masked tokens of each model.

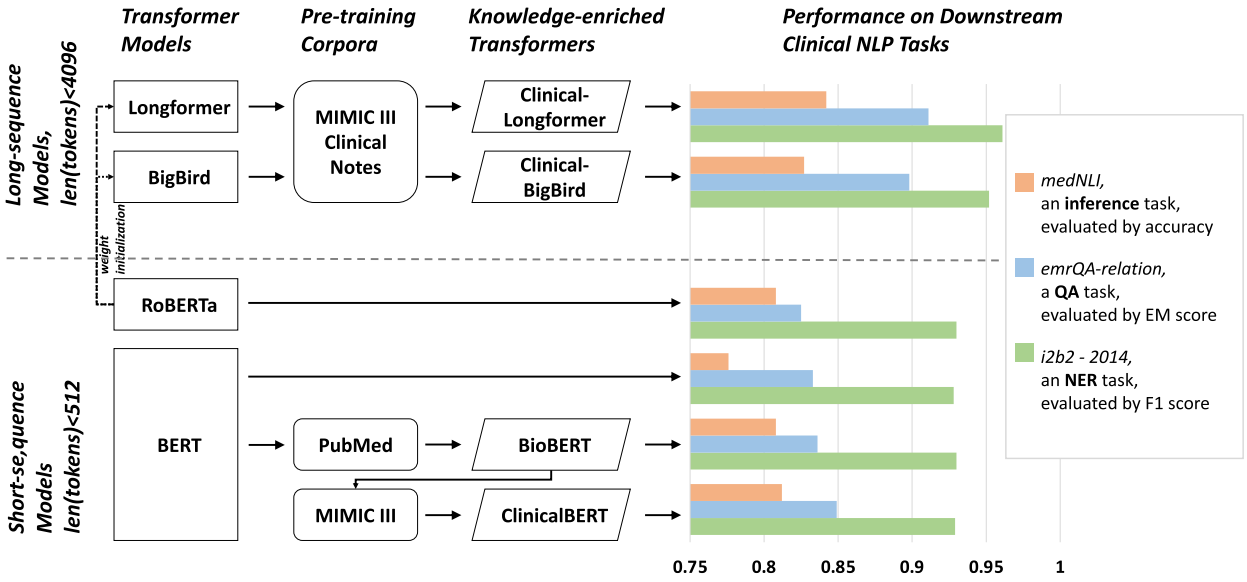

Figure 1. The pipeline for pretraining and fine-tuning transformer-based language models.

Downstream tasks

In this study, we fine-tune the pretrained Clinical-Longformer and Clinical-BigBird on 10 clinical NLP datasets. These 10 NLP datasets broadly cover various NLP tasks, including extractive QA, named entity recognition, natural language inference, and document classification. We rely on these NLP tasks to validate the performance improvement of long-sequence models compared to their short-sequence counterparts. The statistics and descriptions of all datasets can be found in Table 1.

Question answering

QA is a common NLP task that aims to automatically answer questions asked in natural language.[26] In the clinical context, QA systems answer clinicians' questions by understanding the clinical narratives extracted from EHR systems to support decision-making. emrQA[27] is the most frequently used benchmark dataset in clinical QA, which contains more than 400 000 question-answer pairs semi-automatically generated from past Informatics for Integrating Biology and the Bedside (i2b2) challenges. emrQA falls into the category of extractive QA, aiming to identify answer spans from reference texts instead of generating new answers in a word-by-word fashion. Researchers have attempted to solve emrQA tasks by using word embedding models,[28] conditional random fields (CRFs),[29] and transformer-based models,[30] among which transformer-based models performed best. In our experiments, we investigate the performance of our pretrained models using the 3 largest emrQA subsets: *Medication*, *Relation,* and *Heart Disease*. We evaluate QA performance with 2 commonly used metrics: exact match (EM) and F1-score. Exact match evaluates whether entire predicted spans match exactly with the ground-truth annotations. F1-score is a looser metric derived from token-level precision and recall, which measures the overlap between the predictions and the targets. We generate train-dev-test splits by following the instruction of Yue et al.[28] The training set of *relation* and *medication* subsets are randomly under-sampled to reduce training time. Based on their experience, performance was not compromised by under-sampling. Of note, the emrQA dataset has some known issues, for example, incomplete answers, it is template-based, and the annotation were generated semiautomatically.[28] We consider the usage of emrQA as a proof-of-concept experiment to compare the performance of the transformer-based model on the QA task.

Named entity recognition

Named entity recognition is a token-level classification task that seeks to identify the named entities and classify them into predefined categories. This genre of NLP tasks has broad applications in the clinical and biomedical domains, for example, the deidentification of PHI and medical concept extraction from clinical notes. Prior studies have shown that transformer-based models[15] significantly outperformed the models built on pretrained static word embeddings[31] or LSTM networks.[32] We compare our pretrained models using 4 data challenges: (1) i2b2 2006[33] to deidentify PHI from medical discharge notes. (2) i2b2 2010[34] to extract and annotate medical concepts from patient reports, (3) i2b2 2012[35] to identify both clinical concepts and events relevant to the patient's clinical timeline from discharge summaries, and (4) i2b2 2014[36] to identify PHI information from longitudinal clinical narratives. We follow the processing steps of Alsentzer et al,[16] which convert the raw data from all 4 tasks to the inside-outside-beginning tagging format proposed by Ramshaw et al,[37] and then create train-dev-test splits. We evaluate the model performance with an F1 score similar to QA tasks.

Document classification

Document classification is one of the most common NLP tasks, where a sentence or document is assigned to one or more classes or categories. In the clinical domain, document classification can be used to identify the onset of a particular disease process or predict patient prognosis using entire clinical notes. We use the following 3 document classification datasets to evaluate the pretrained models from different perspectives.

MIMIC-AKI[38,39] MIMIC-AKI is a binary classification task, where we aim to predict the possibility of acute kidney injury (AKI) for critically ill patients using the clinical notes within the first 24 hours following ICU admission. We follow Li et al[38] to extract the cohort from MIMIC-III. We evaluate the model performance using AUC (area under the receiver operating characteristic curve) and F1 score.

OpenI[40] OpenI is a publicly available chest X-ray (CXR) dataset collected by Indiana University. The dataset provides around 4000 radiology reports and their associated human-annotated Medical Subject Headings (MeSH) terms. In our experiments, the task is to detect the presence of the annotated thoracic findings from CXR reports, which is considered a multilabel classification task. Given the small sample size, we will only use OpenI as the testing set. The pretrained language models are fine-tuned using MIMIC-CXR,[41] another publicly available CXR dataset that contains more than 200 000 CXR reports. Unlike openI, the ground-truth labels for MIMIC-CXR were automatically generated using NLP approaches. The overlapping findings between the 2 CXR data sources are *Car-*

Table 1. Description and statistics of downstream clinical NLP tasks

Dataset	Task	Source	Sample size	Avg. seq. length	Max seq. length
MedNLI	Inference	MIMIC	14 049	39	409
i2b2 2006	NER	i2b2	66 034	867	3986
i2b2 2010	NER	i2b2	43 947	1459	6052
i2b2 2012	NER	i2b2	13 108	794	2900
i2b2 2014	NER	i2b2	83 466	5134	14 370
emrQA-Relation	QA	i2b2	255 908	1880	6109
emrQA-Medication	QA	i2b2	141 243	1460	6050
emrQA-HeartDisease	QA	i2b2	30 731	5293	14 060
openI	Multilabel Classif.	IndianaU	3684	70	294
MIMIC-CXR	Multilabel Classif.	MIMIC-CXR	222 713	119	874
MIMIC-AKI	Binary Classif.	MIMIC	16 536	1463	20 857

diomegaly, *Edema*, *Consolidation*, *Pneumonia*, *Atelectasis*, *Pneumothorax*, and *Pleural Effusion*. We report the sample number weighted average of the area under the receiver operating characteristic curve (AUC) as proposed and used in Li et al[42]and Wang et al.[43]

MedNLI[44] Natural language inference (NLI) is the task of determining the relationship between sentence pairs. MedNLI is a collection of sentence pairs extracted from the MIMIC-III[24] and annotated by 2 board-certified radiologists. The relationship between the premise sentence and the hypothesis sentence could be entailment, contradiction, or neutral. Transformer-based models process NLI tasks also as document classification by merging the sentence pair and placing a delimiter token right after the end of the first sentence. We follow the original splits as Romanov et al[44] and use accuracy to evaluate the performance.

Baseline models and comparisons

Both Clinical-Longformer and Clinical-BigBird are compared with the short-sequence models, including BERT, ClinicalBERT, RoBERTa, and BioBERT. We do not include the static word embedding models, for example, Word2Vec and FastText, in the comparisons, because those models yield less competitive performance compared to the transformer-based models and cannot easily handle token-level classification tasks. The BERT[2] model is the first-of-its-kind transformer architecture that achieved state-of-the-art results on eleven NLP tasks. Both masked language modeling and next-sentence prediction were used to learn contextualized word representation from BooksCorpus and English Wikipedia in the pretraining stage. BioBERT[15] is the first biomedical domain-specific BERT variant pretrained from PubMed abstracts and PMC full-text articles. The weights of BioBERT were initialized from BERT. BioBERT yielded optimal performance in biomedical QA, NER, and relation extraction tasks. ClinicalBERT,[16] initialized from BioBERT, was further pretrained using clinical notes also extracted from MIMIC-III. ClinicalBERT boosted the performance for MedNLI and 4 i2b2 NER tasks that are also included in our study. BioBERT and ClinicalBERT use the next sentence prediction and masked language modeling as pretraining strategies. RoBERTa[3] is an improved variant of BERT model, which is trained with larger corpus, bigger batch size, and gets rid of the next sentence prediction objectives. Both Longformer and BigBird initialize their training weights from RoBERTa checkpoint. We also try hierarchical transformers[45] in the experiment of MIMIC-AKI. The hierarchical transformer model uses the BERT model to learn outputs from each small chunk of text. Then, the outputs of small chunks are fed into the recurrent neural network. Given that the hierarchical transformer model is not explicitly developed for clinical NLP, we load the weights of ClinicalBERT to initialize the BERT layers.

Experimental setup

For the token-level classification, including QA and NER, a classification head is added to the output of each token obtained from the transformer-based architecture. The sequences are split into chunks in the length of 4096 for Clinical-Longformer and Clinical-BigBird, and 512 for all the other 3 baseline models. One thousand and twenty-four strides are taken between the chunks of long-sequence models; 128 strides are taken between the chunks of short-sequence models.

For the document classification tasks, the predicted outcomes are derived from the [CLS] token added to the beginning of each sentence or document. The maximum sequence lengths of OpenI and MedNLI are less than 512 tokens. Therefore, no truncation or sliding window approaches are needed for these 2 datasets. In MIMIC-AKI, given that some clinical notes are extremely long, which may even exceed the length limits of all models, we first truncate each document to the first 4096 tokens, which meets the length limits of Clinical-Longformer and Clinical-BigBird. The predicted outcomes are directly derived from the [CLS] output when using both long-sequence models. When dealing with short-sequence models, the documents are further segmented into snippets of 512 tokens in order to accommodate for the length requirement of short-sequence models. A pooling strategy, which was introduced by Huang et al[10] to predict ICU readmission from discharge summaries, is applied to aggregate the probability outputs from short snippets. The probability of AKI onset for a patient with n short snippets is computed by: $P_{\text{AKI}} = \frac{[\max_{i\in n} p_i]^n + \frac{n}{2} * [\frac{1}{n}\sum_i^n p_i]^n}{1+\frac{n}{2}}$, where p_i is the probability output of the ith snippet from the short-sequence model. Our preliminary experiments show that this pooling strategy slightly outperforms the maximum pooling method.

We conduct our experiments using four 32GB GPUs. We maximize the batch size for each experiment given the memory limits of GPUs to save training time. The batch size during training is 16 for Clinical-Longformer, 12 for Clinical-BigBird, and 64 for all other models. Batch sizes are doubled when evaluating or testing. Half precision is applied to both Clinical-Longformer and Clinical-BigBird. We try learning rates: {1e−5, 2e−5, and 5e−5} for the experiments of each model on each task. We fine-tune 6 epochs for each set-up. All experiments converge within 6 epochs. The best-performing model parameters are determined by the performance of the development split. The experiments are implemented with python 3.8.0, PyTorch 1.9.0, and Transformer 4.9.0. The versions and downloadable links for all models can be found in Supplementary Table S1.

RESULTS AND DISCUSSION

The evaluation of pretraining can be found in Table 2. The results demonstrate that Clinical-Longformer and Clinical-Bigbird can learn more useful contextualized relationships from clinical notes in the pretraining when compared to other baseline models, which provides the foundation for performance improvement in downstream tasks. BERT, BioBERT, and RoBERTa which are not pretrained using clinical notes, yield very poor perplexity scores and masked language modeling (MLM) accuracies. This confirms that pretraining using domain-specific corpus is essential for learning the domain-specific contextualized relationships. We also visualize an

Table 2. The evaluation of transformer-based models after language modeling (LM) pretraining

Pretrained models	Perplexity score	MLM accuracy
BERT	52 807.11	0.633
BioBERT	131 176.11	0.001
ClinicalBERT	8.67	0.803
RoBERTa	1378.71	0.693
Clinical-Longformer	<u>1.61</u>	**0.940**
Clinical-BigBird	**1.41**	<u>0.936</u>

Note: The best scores are in bold, and the second-best scores are underlined.

Table 3. The performance of transformer-based pretrained models on question answering tasks

Pretrained models	emrQA-medication		emrQA-relation		emrQA-heart disease	
metrics	EM	F1	EM	F1	EM	F1
BERT	0.240	0.675	0.833	0.924	0.650	0.698
BioBERT	0.247	0.700	0.836	0.926	0.647	0.702
ClinicalBERT	0.297	0.698	0.849	0.929	0.666	0.711
RoBERTa	0.280	0.706	0.825	0.917	0.655	0.682
Clinical-Longformer	**0.302**	**0.716**	**0.911**	**0.948**	**0.698**	**0.734**
Clinical-BigBird	0.300	0.715	0.898	0.944	0.664	0.711

Note: The best scores are in bold, and the second-best scores are underlined.

Table 4. The performance of transformer-based pretrained models on NER tasks

Pretrained models *metrics*	i2b2 2006 F1	i2b2 2010 F1	i2b2 2012 F1	i2b2 2014 F1
BERT	0.939	0.835	0.759	0.928
BioBERT	0.948	0.865	0.789	0.930
ClinicalBERT	0.951	0.861	0.773	0.929
RoBERTa	0.956	0.851	0.767	0.930
Clinical-Longformer	**0.974**	**0.887**	**0.800**	**0.961**
Clinical-BigBird	0.967	0.872	0.787	0.952

Note: The best scores are in bold, and the second-best scores are underlined.

Table 5. The performance of transformer-based models on document classification tasks

Pretrained models	OpenI	MIMIC-AKI		medNLI
metrics	Accuracy	AUC	F1	Accuracy
BERT	0.952	0.545	0.296	0.776
BioBERT	0.954	0.717	0.372	0.808
ClinicalBERT	0.967	0.747	0.468	0.812
RoBERTa	0.963	0.708	0.358	0.808
Hierarchical transformer	–	0.726	0.462	–
Clinical-Longformer	**0.977**	**0.762**	**0.484**	**0.842**
Clinical-BigBird	0.972	0.755	0.480	0.827

Note: The best scores are in bold, and the second-best scores are underlined.

example in Supplementary Figure S1. When [Stroke] is replaced with a mask token, Clinical-Longformer can infer this word from [infarct], [hemorrhagic], [epilepticus], and [hemorrhage], which are more than 1000 tokens away from the mask token. This example demonstrates that our models can learn long-term dependencies from clinical narratives.

Full results for QA, NER, and classification tasks are presented in Tables 3, 4, and 5, respectively (for full results with variance measurements, please see Supplementary Tables S2–S4). In QA tasks, both Clinical-Longformer and Clinical-BigBird outperform the short-sequence transformer models by around 2% across all 3 emrQA subsets when evaluated by F1 score. When considering the stricter, EM metric, Clinical-Longformer, and Clinical-BigBird improve ~5% on the relations subset but yield similar results to ClinicalBERT in the other 2 subsets. In NER tasks, Clinical-Longformer consistently leads the short-sequence transformers by more than 2% in all 4 i2b2 datasets. Clinical-BigBird also performs better than ClinicalBERT and BioBERT in all NER experiments. In document classification tasks, our 2 long-sequence transformers achieve superior results compared to prior models on OpenI, MIMIC-AKI, and medNLI tasks.

We observe that Clinical-Longformer and Clinical-BigBird not only improve the performance of long sequences tasks but also short sequences. The maximum sequences of MedNLI and OpenI are smaller than 512 tokens, which can be entirely fed into the BERT-like models. However, the long-sequence models still achieve better results. We attribute these improvements to the pretraining stages of Clinical-Longformer and Clinical-BigBird, where the language models can learn more long-term dependencies by extending the sequence length limit, thereby learning a richer contextualization of clinical concepts. We find more significant gains, however, when applying our 2 long-sequence models to the datasets with longer sequences. For example, the performance improvement is most dramatic on the i2b2 2014 dataset, which has the largest average sequence length in all 4 NER tasks (almost twice the other 3 subsets). Likewise, Clinical-Longformer more strongly improves the F1 score of the *heart disease* subset from emrQA. This suggests that Clinical-Longformer and Clinical-BigBird are also better at modeling long-term dependencies in downstream tasks. Moreover, in i2b2 2006 dataset, the models achieve superior results in identifying the PHI information from the clinical notes. However, all PHI placeholders are completely removed in the preprocessing step of pretraining. This confirms that the language models can be generalized to new tokens in downstream tasks that are unseen in pretraining stage. Finally, we also find that Clinical-Longformer yields slightly better results when compared to Clinical-BigBird, although the differences in most experiments are not statistically significant. Given that Clinical-BigBird also requires more fine-tuning time and memory costs, we recommend that future investigators apply our Clinical-Longformer checkpoint to their own tasks when resources are limited.

Our study has several limitations. Firstly, we only apply Longformer and BigBird to large-scale clinical corpus. In future iterations, we plan to release more pretrained models for long sequences enriched with other biomedical corpora, for example, PubMed and PMC publications. Also, we only pretrain the base cased version of Clinical-Longformer and Clinical-BigBird. We will publish the uncased and large version at the next step. Secondly, another recent approach developed to address the memory problem of long sequences is simplifying or compressing the transformer architecture. In future work, we will compare this genre of transformers, for example, TinyBERT,[46] to our current long-sequence models. Thirdly, we do not integrate Clinical-Longformer or Clinical-BigBird into an encoder-decoder framework due to the memory limits of our GPU

cards. Therefore, experiments on generative tasks like text generation or document summarization are not included in this study. We intend to incorporate these tasks into future versions of these models as our computational capability evolves. Fourthly, the emrQA was annotated in a semiautomatic way without expert calibration. There are incorrect NER labels as mentioned in Yue et al.[28] We will conduct the experiments on a large-scale human-annotated NER dataset should there be any availability. Finally, the vocabularies of Clinical-Longformer and Clinical-BigBird are inherited from the 5000 subword units used in the RoBERTa[3] model that was developed for nonclinical corpus. We have no idea if other types of tokenizers or a clinical-adaptive vocabulary can boost the performance. Therefore, we will examine more combinations in future studies.

CONCLUSION

In this study, we introduce Clinical-Longformer and Clinical-BigBird, 2 pretrained language models designed specifically for long clinical text NLP tasks. We compare these 2 models with the BERT-variant short-sequence transformer-based models, for example, ClinicalBERT, in named entity recognition, QA, and document classification tasks. Results demonstrate that Clinical-Longformer and Clinical-BigBird achieve better results on both long- and short-sequence benchmark datasets. Future studies will investigate the generalizability of our proposed models to clinical text generation and summarization tasks, and the comparison with other modeling approaches that are also developed to solve the memory consumption of long text.

FUNDING

The National Institutes of Health grant number U01TR003528 and R01LM013337.

AUTHOR CONTRIBUTIONS

YLi, RMW, FSA, and YLuo conceived of the presented idea. YLi and HW carried out the experiments. YLi, RMW, FSA, and YLuo contributed to interpreting the results. YLi wrote the manuscript in consultation with YLuo. All authors provided critical feedback and helped shape the research, analysis, and manuscript.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

CONFLICT OF INTEREST STATEMENT

None declared.

DATA AVAILAILITY

The benchmark datasets are derived from multiple publicly available datasets, including MIMIC III from https://physionet.org/content/mimiciii/1.4/; MIMIC-CXR from https://physionet.org/content/mimic-cxr/2.0.0/; i2b2 from https://portal.dbmi.hms.harvard.edu/; openI from https://openi.nlm.nih.gov/; and MedNLI from https://physionet.org/content/mednli/1.0.0/. To officially gain access, the authors should apply and sign data user agreement with the data owner. We provide codes to preprocess and generate splits at: https://github.com/luoyuanlab/Clinical-Longformer.

REFERENCES

1. Brown T, *et al.* Language models are few-shot learners. *Adv Neural Inform Process Syst* 2020; 33: 1877–901.
2. Devlin J, *et al.* Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers).Minneapolis, MN: Association for Computational Linguistics; 2019: 4171–86.
3. Liu Y, *et al.* Roberta: a robustly optimized Bert pretraining approach. arXiv preprint arXiv:1907.11692, 2019.
4. Yao L, Jin Z, Mao C, *et al.* Traditional Chinese medicine clinical records classification with BERT and domain specific corpora. *J Am Med Inform Assoc* 2019; 26 (12): 1632–6.
5. Zhang Z, Liu J, Razavian N. BERT-XML: large scale automated ICD coding using BERT pretraining. In: Proceedings of the 3rd Clinical Natural Language Processing Workshop; 2020.
6. Liu W, *et al.* K-bert: Enabling language representation with knowledge graph. *Proc Conf AAAI Artif Intell* 2020; 34 (3).
7. Wen A, Elwazir MY, Moon S, *et al.* Adapting and evaluating a deep learning language model for clinical why-question answering. *JAMIA Open* 2020; 3 (1): 16–20.
8. Vaswani A, *et al.* Attention is all you need. *Adv Neural Inform Process Syst* 2017; 30.
9. Gao S, Alawad M, Young MT, *et al.* Limitations of transformers on clinical text classification. *IEEE J Biomed Health Inform* 2021; 25 (9): 3596–607.
10. Huang K, Singh A, Chen S, *et al.* Clinical XLNet: modeling sequential clinical notes and predicting prolonged mechanical ventilation. In: Proceedings of the 3rd Clinical Natural Language Processing Workshop; 2020: 94–100.
11. Mahbub M, Srinivasan S, Danciu I, *et al.* Unstructured clinical notes within the 24 hours since admission predict short, mid & long-term mortality in adult ICU patients. *PLoS One* 2022; 17 (1): e0262182.
12. Ainslie J, *et al.* ETC: Encoding long and structured inputs in transformers. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP); 2020.
13. Beltagy I, Peters ME, Cohan A. Longformer: the long-document transformer. arXiv preprint arXiv:2004.05150, 2020.
14. Zaheer M, *et al.* Big bird: transformers for longer sequences. *Adv Neural Inform Process Syst* 2020; 33: 17283–97.
15. Lee J, Yoon W, Kim S, *et al.* BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 2020; 36 (4): 1234–40.
16. Alsentzer E, Murphy J, Boag W, *et al.* Publicly available clinical BERT embeddings. In: Proceedings of the 2nd Clinical Natural Language Processing Workshop; 2019: 72–8.
17. Smit A, Jain S, Rajpurkar P, Pareek A, Ng AY, Lungren M. Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP); 2020: 1500–19.
18. He Y, *et al.* Infusing disease knowledge into BERT for health question answering, medical inference and disease name recognition. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP); 2020.
19. Michalopoulos G, *et al.* UmlsBERT: clinical domain knowledge augmentation of contextual embeddings using the unified medical language system metathesaurus. In: Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: human language technologies. 2021.
20. Zhou S, Wang N, Wang L, Liu H, Zhang R. CancerBERT: a cancer domain-specific language model for extracting breast cancer phenotypes

from electronic health records. *J Am Med Inform Assoc* 2022;29 (7):1208–16.
21. Agrawal MN, Lang H, Offin M, Gazit L, Sontag D. Leveraging time irreversibility with order-contrastive pre-training. In: International Conference on Artificial Intelligence and Statistics; PMLR; 2022: 2330–53.
22. Dai Z, Yang Z, Yang Y, Carbonell JG, Le Q, Salakhutdinov R. Transformer-XL: attentive language models beyond a fixed-length context. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics; 2019: 2978–88.
23. Kitaev N, Kaiser L, Levskaya A. Reformer: the efficient transformer. In: 8th International Conference on Learning Representations, ICLR 2020; April 26–30, 2020; Addis Ababa, Ethiopia. OpenReview.net.
24. Johnson AE, Pollard TJ, Shen L, *et al.* MIMIC-III, a freely accessible critical care database. *Sci Data* 2016; 3 (1): 1–9.
25. Wang C, Cho K, Gu J. Neural machine translation with byte-level subwords. In: Proceedings of the AAAI conference on artificial intelligence; 2020; New York.
26. Cimiano P, Unger C, McCrae J. Ontology-based *Interpretation of Natural Language*. vol. 7 (2) *Synthesis Lectures on Human Language Technologies*. Morgan & Claypool Publishers; 2014: 1–178.
27. Pampari A, Raghavan P, Liang J, Peng J. emrQA: A large corpus for question answering on electronic medical records. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing; 2018: 2357–68.
28. Yue X, Jimenez B, Sun H. Clinical reading comprehension: a thorough analysis of the emrQA dataset. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL'20); 2020.
29. Kang M, Han M, Hwang SJ. Neural mask generator: learning to generate adaptive word maskings for language model adaptation. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP); 2020: 6102–20.
30. Soni S, Roberts K. Evaluation of dataset selection for pre-training and fine-tuning transformer language models for clinical question answering. In: Proceedings of the 12th language resources and evaluation conference; 2020.
31. Wang X, Zhang Y, Ren X, *et al.* Cross-type biomedical named entity recognition with deep multi-task learning. *Bioinformatics* 2019; 35 (10): 1745–52.
32. Yoon W, So CH, Lee J, *et al.* Collabonet: collaboration of deep neural networks for biomedical named entity recognition. *BMC Bioinformatics* 2019; 20 (S10): 55–65.
33. Uzuner Ö, Luo Y, Szolovits P. Evaluating the state-of-the-art in automatic de-identification. *J Am Med Inform Assoc* 2007; 14 (5): 550–63.
34. Uzuner Ö, South BR, Shen S, *et al.* VA challenge on concepts, assertions, and relations in clinical text. *J Am Med Inform Assoc* 2011; 18 (5): 552–6.
35. Sun W, Rumshisky A, Uzuner O. Evaluating temporal relations in clinical text: 2012 i2b2 challenge. *J Am Med Inform Assoc* 2013; 20 (5): 806–13.
36. Stubbs A, Uzuner Ö. Annotating longitudinal clinical narratives for de-identification: the 2014 i2b2/UTHealth corpus. *J Biomed Informatics* 2015; 58: S20–S29.
37. Ramshaw LA, Marcus MP. Text chunking using transformation-based learning. In: *Natural language processing using very large corpora*; 1999; Springer. p. 157–176.
38. Li Y, *et al.* Early prediction of acute kidney injury in critical care setting using clinical notes. In: 2018 IEEE international conference on bioinformatics and biomedicine (BIBM); 2018; IEEE; Madrid, Spain.
39. Sun M, *et al.* Early prediction of acute kidney injury in critical care setting using clinical notes and structured multivariate physiological measurements. *MedInfo* 2019; 264: 368–72.
40. Demner-Fushman D, Antani S, Simpson M, *et al.* Design and development of a multimodal biomedical information retrieval system. *J Comput Sci Eng* 2012; 6 (2): 168–77.
41. Johnson AEW, Pollard TJ, Berkowitz SJ, *et al.* MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Sci Data* 2019; 6 (1): 1–8.
42. Li Y, Wang H, Luo Y. A comparison of pre-trained vision-and-language models for multimodal representation learning across medical images and reports. In: 2020 IEEE international conference on bioinformatics and biomedicine (BIBM); 2020; IEEE.
43. Wang X, *et al.* Tienet: text-image embedding network for common thorax disease classification and reporting in chest X-rays. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2018; Salt Lake City.
44. Romanov A, Shivade C. Lessons from natural language inference in the clinical domain. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing; 2018: 1586–96.
45. Pappagari R, *et al.* Hierarchical transformers for long document classification. In: 2019 IEEE automatic speech recognition and understanding workshop (ASRU); 2019; IEEE; Sentosa, Singapore.
46. Jiao X, Yin Y, Shang L, *et al.* TinyBERT: distilling BERT for natural language understanding. In: Findings of the Association for Computational Linguistics: EMNLP 2020; 2020: 4163–74.