ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission

Kexin Huang
Health Data Science, Harvard T.H. Chan School of Public Health

Jaan Altosaar
Department of Physics, Princeton University

Rajesh Ranganath
Courant Institute of Mathematical Science, New York University

Abstract

Clinical notes contain information about patients beyond structured data such as lab values or medications. However, clinical notes have been underused relative to structured data, because notes are high-dimensional and sparse. We aim to develop and evaluate a continuous representation of clinical notes. Given this representation, our goal is to predict 30-day hospital readmission at various timepoints of admission, including early stages and at discharge. We apply bidirectional encoder representations from transformers (BERT) to clinical text. Publicly-released BERT parameters are trained on standard corpora such as Wikipedia and BookCorpus, which differ from clinical text. We therefore pre-train BERT using clinical notes and fine-tune the network for the task of predicting hospital readmission. This defines ClinicalBERT. ClinicalBERT uncovers high-quality relationships between medical concepts, as judged by physicians. ClinicalBERT outperforms various baselines on 30-day hospital readmission prediction using both discharge summaries and the first few days of notes in the intensive care unit on various clinically-motivated metrics. The attention weights of ClinicalBERT can also be used to interpret predictions. To facilitate research, we open-source model parameters, and scripts for training and evaluation. ClinicalBERT is a flexible framework to represent clinical notes. It improves on previous clinical text processing methods and with little engineering can be adapted to other clinical predictive tasks.

ACM Reference Format:

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2020. ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission. In *CHIL '20: ACM Conference on Health, Inference, and Learning; Workshop Track. April 02–04, 2020, Toronto, ON*. ACM, New York, NY, USA, 9 pages.

1 Introduction

An electronic health record (EHR) stores patient information; it can save money, time, and lives [21]. Data is added to an EHR daily, so analyses may benefit from machine learning. Machine learning techniques leverage structured features in EHR data, such as lab results or electrocardiography measurements, to uncover patterns and improve predictions [30, 36, 37]. However, unstructured, high-dimensional, and sparse information such as clinical notes are difficult to use in clinical machine learning models. Our goal is to create a framework for modeling clinical notes that can uncover clinical insights and make medical predictions.

Clinical notes contain significant clinical value [5, 35, 18, 34]. A patient might be associated with hundreds of notes within a stay and over their history of admissions. Compared to structured features, clinical notes provide a richer picture of the patient since they describe symptoms, reasons for diagnoses, radiology results, daily activities, and patient history. Consider clinicians working in the intensive care unit, who need to make decisions under time constraints. Making accurate clinical predictions may require reading a large volume of clinical notes. This can add to a doctor's workload, so tools that make accurate predictions based on clinical notes might be useful in practice.

Hospital readmission lowers patients' quality of life and wastes money [2, 40]. One estimate puts the financial burden of readmission at $17.9 billion and the fraction of avoidable admissions at 76% [4]. Accurately predicting readmission has clinical significance, as it may improve efficiency and reduce the burden on intensive care unit doctors. We develop a discharge support model, ClinicalBERT, that processes patient notes and dynamically assigns a risk score of whether the patient will be readmitted within 30 days (Figure 1). As physicians and nurses write notes about a patient, ClinicalBERT processes the notes and updates the risk score of readmission. This score can inform provider decisions, such as whether to intervene. Besides readmission, ClinicalBERT can be adapted to other tasks such as diagnosis prediction, mortality risk estimation, or length-of-stay assessment.

1.1 Background

Electronic health records are useful for risk prediction [13]. Clinical notes in such electronic health records use abbreviations, jargon, and have an unusual grammatical structure. Building models that learn useful representations of clinical text is a challenge [9]. Bag-of-words assumptions have been used to model clinical text [38], in addition to log-bilinear word embedding models such as Word2Vec [20, 23]. The latter word embedding models learn representations of clinical text using local contexts of words. But clinical notes are long and their words are interdependent [39], so these methods cannot capture the long-range dependencies needed to capture clinical meaning.

Natural language processing methods where representations include global, long-range information can yield boosts in performance on clinical tasks [24, 25, 11]. Modeling clinical notes requires capturing interactions between distant words. The need to model this long-range structure makes clinical notes suitable for contextual representations like bidirectional encoder representations from transformers (BERT) [11]. Lee et al. [17] apply BERT to biomedical literature, and [31] use BERT to enhance clinical concept extraction. Concurrent to our work, Alsentzer et al. [1] also apply BERT to clinical notes; we

CHIL '20 Workshop, April 02–04, 2020, Toronto, ON

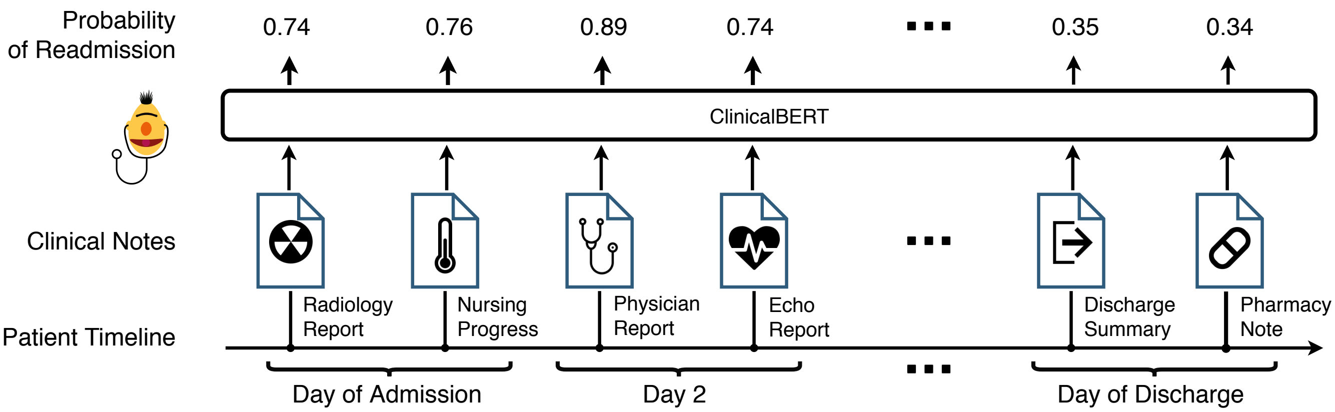

Figure 1: ClinicalBERT learns deep representations of clinical notes that are useful for tasks such as readmission prediction. In this example, care providers add notes to an electronic health record during a patient's admission, and the model dynamically updates the patient's risk of being readmitted within a 30-day window.

evaluate and adapt ClinicalBERT to the clinical task of readmission and pre-train on longer sequence lengths.

Methods to evaluate models of clinical notes are also relevant to ClinicalBERT. Wang et al. [34] and Chiu et al. [10] evaluate the quality of biomedical embeddings by computing correlations between doctor-rated relationships and embedding similarity scores. We adopt similar evaluation techniques in our work.

Good representations of clinical text require good performance on downstream tasks. We use 30-day hospital readmission prediction as a case study since it is of clinical importance. We refer readers to Futoma et al. [12] for comparisons of traditional machine learning methods such as random forests and neural networks on hospital readmission tasks. Work in this area has focused on integrating a multitude of covariates about a patient into a model [7]. Caruana et al. [8] develop an interpretable model for readmission prediction based on generalized additive models and highlight the need for intelligible clinical predictions. Rajkomar et al. [26] predict readmission using a standard ontology from notes alongside structured information. Much of this previous work uses information at discharge, whereas ClinicalBERT can predict readmission during a patient's stay.

1.2 Significance

ClinicalBERT improves readmission prediction over methods that center on discharge summaries. Making a prediction using a discharge summary at the end of a stay means that there are fewer opportunities to reduce the chance of readmission. To build a clinically-relevant model, we define a task of predicting readmission at any timepoint since a patient was admitted. To evaluate models on readmission prediction, we define a metric motivated by a clinical challenge. Medicine suffers from alarm fatigue [28, 3]. This means useful classification rules for medicine need to have high positive predictive value (precision). We evaluate model performance at a fixed positive predictive value. We show that ClinicalBERT has the highest recall compared to popular methods for representing clinical notes. ClinicalBERT can be readily applied to other tasks such as mortality prediction and disease prediction. In addition, ClinicalBERT attention weights can be visualized to understand which elements of clinical notes are relevant to a prediction.

ClinicalBERT is BERT [11] specialized to clinical notes. Clinical notes are lengthy and numerous, and the computationally-efficient architecture of BERT can model long-term dependencies. Compared to two popular models of clinical text, Word2Vec and FastText, ClinicalBERT more accurately captures clinical word similarity. We describe one way to scale up ClinicalBERT to handle large collections of clinical notes for clinical prediction tasks. In a case study of hospital readmission prediction, ClinicalBERT outperforms competitive deep language models. We open source ClinicalBERT[1] pre-training and readmission model parameters along with scripts to reproduce results and apply the model to new tasks.

2 Methods

ClinicalBERT learns deep representations of clinical text. These representations can uncover clinical insights (such as predictions of disease), find relationships between treatments and outcomes, or create summaries of corpora. ClinicalBERT is an application of the BERT model [11] to clinical corpora to address the challenges of clinical text. Representations are learned using medical notes and further processed for clinical tasks; we demonstrate ClinicalBERT on the task of hospital readmission prediction.

2.1 BERT Model

BERT is a deep neural network that uses the transformer encoder architecture [33] to learn embeddings for text. We omit a detailed description of the architecture; it is described in [33]. The transformer encoder architecture is based on a self-attention mechanism. The pre-training objective function for the model is defined by two unsupervised tasks: masked language modeling and next sentence prediction. The text embeddings and model parameters are fit using stochastic optimization. For downstream tasks, the fine-tuning phase is problem-specific; we describe a fine-tuning task specific to clinical text.

[1] https://github.com/kexinhuang12345/clinicalBERT

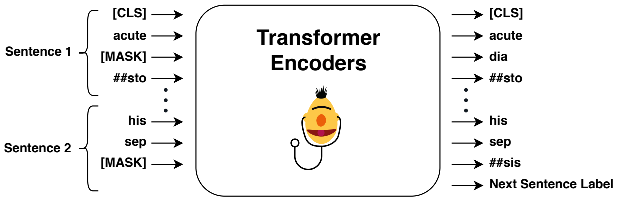

Figure 2: ClinicalBERT learns deep representations of clinical text using two unsupervised language modeling tasks: masked language modeling and next sentence prediction. In masked language modeling, a fraction of input tokens are held out for prediction; in next sentence prediction, ClinicalBERT predicts whether two input sentences are consecutive.

2.2 Clinical Text Embedding

A clinical note input to ClinicalBERT is represented as a collection of tokens. These tokens are subword units extracted from text in a preprocessing step [29]. In ClinicalBERT, a token in a clinical note is represented as a sum of the token embedding, a learned segment embedding, and a position embedding. When multiple sequences of tokens are fed to ClinicalBERT, the segment embedding identifies which sequence a token is associated with. The position embedding of a token is a learned set of parameters corresponding to the token's position in the input sequence (position embeddings are shared across tokens). A classification token [CLS] is inserted in front of every sequence of input tokens for use in classification tasks.

2.3 Self-Attention Mechanism

The attention function is computed on an input sequence using the embeddings associated with the input tokens. The attention function takes as input a set of queries, keys, and values. To construct the queries, keys, and values, every input embedding is multiplied by learned sets of weights (it is called 'self' attention because the values are the same as the keys and queries). For a single query, the output of the attention function is a weighted combination of values. The query and a key determine the weight for a value. Denote a set of queries, keys, and values by Q, K, and V. The attention function is

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d}}V), \tag{1}$$

where d is the dimensionality of the queries, keys, and values. This function can be computed efficiently and can capture long-range interactions between any two elements of the input sequence [33]. The length and complex patterns in clinical notes makes the transformer architecture with self-attention a good choice. (We later describe how this attention mechanism can allow interpretation of ClinicalBERT predictions.)

2.4 Pre-training ClinicalBERT

The quality of learned representations of text depends on the text the model was trained on. BERT is trained on BooksCorpus and Wikipedia. But these datasets are distinct from clinical notes, as jargon and abbreviations prevail: clinical notes have different syntax and grammar than books or encyclopedias. These differences make clinical notes hard to understand without expertise. ClinicalBERT is pre-trained on clinical notes as follows.

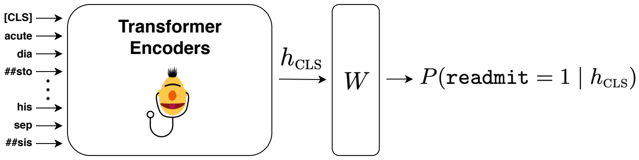

Figure 3: ClinicalBERT models clinical notes and can be readily adapted to clinical tasks such as predicting 30-day readmission. The model is fed a patient's clinical notes, and the patient's risk of readmission within a 30-day window is predicted using a linear layer applied to the classification representation $h_{[\text{CLS}]}$ learned by ClinicalBERT. This fine-tuning task is described in Equation (2)

ClinicalBERT uses the same pre-training tasks as [11]. Masked language modeling means masking some input tokens and training the model to predict the masked tokens. In next sentence prediction, two sentences are fed to the model. The model predicts whether these sentences are consecutive. The pre-training objective function is the sum of the log-likelihood of the predicted masked tokens and the log-likelihood of the binary variable indicating whether two sentences are consecutive.

2.5 Fine-tuning ClinicalBERT

After pre-training, ClinicalBERT is fine-tuned on a clinical task: readmission prediction. Let readmit be a binary indicator of readmission of a patient in the next 30 days. Given clinical notes as input, the output of ClinicalBERT is used to predict the probability of readmission:

$$P(\texttt{readmit} = 1 | h_{[\text{CLS}]}) = \sigma(W h_{[\text{CLS}]}) \tag{2}$$

where σ is the sigmoid function, $h_{[\text{CLS}]}$ is the output of the model corresponding to the classification token, and W is a parameter matrix. The model parameters are fine-tuned to maximize the log-likelihood of this binary classifier.

3 Empirical Study

3.1 Data

We use the Medical Information Mart for Intensive Care III (MIMIC-III) dataset [15]. MIMIC-III consists of the electronic health records of 58,976 unique hospital admissions from 38,597 patients in the intensive care unit of the Beth Israel Deaconess Medical Center between 2001 and 2012. There are 2,083,180 de-identified notes associated with the admissions. Preprocessing of the clinical notes is described in S2. If text that exists in the test set of the fine-tuning task is used for pre-training, then training and test metrics will not be independent. To avoid this, admissions are split into five folds for independent runs, with four folds for pre-training (and training during fine-tuning) and the fifth for testing during fine-tuning.

3.2 Empirical Study I: Language Modeling and Clinical Word Similarity

We developed ClinicalBERT, a model of clinical notes whose representations can be used for clinical tasks. Before evaluating its performance as a model of readmission, we study its performance in two experiments. First, we find that ClinicalBERT outperforms BERT

Table 1: ClinicalBERT improves over BERT on clinical language modeling. We report the five-fold average accuracy of masked language modeling (predicting held-out tokens) and next sentence prediction (a binary prediction of whether two sentences are consecutive), on the MIMIC-III corpus of clinical notes.

Model	Language modeling	Next sentence prediction
ClinicalBERT	0.857 ± 0.002	0.994 ± 0.003
BERT	0.495 ± 0.007	0.539 ± 0.006

in clinical language modeling. Then we compare ClinicalBERT to popular word embedding models using a clinical word similarity task. The relationships between medical concepts learned by ClinicalBERT correlate with human evaluations of similarity.

3.2.1 Clinical Language Modeling. We report the five-fold average accuracy of the masked language modeling and next sentence prediction tasks on the MIMIC-III data in Table 1. BERT underperforms, as it was not trained on clinical text, highlighting the need for building models tailored to clinical data such as ClinicalBERT.

3.2.2 Qualitative Analysis. We test ClinicalBERT on data collected to assess medical term similarity [22]. The data is 30 pairs of medical terms whose similarity is rated by physicians. To compute an embedding for a medical term, ClinicalBERT is fed a sequence of tokens corresponding to the term. Following [11], the sum of the last four hidden states of ClinicalBERT encoders is used to represent each medical term. Medical terms vary in length, so the average is computed over the hidden states of subword units. This results in a fixed 768-dimensional vector for each medical term. We visualize the similarity of medical terms using dimensionality reduction [19], and display a cluster heart-related concepts in Figure 4. Heart-related concepts such as myocardial infarction, atrial fibrillation, and myocardium are close together; renal failure and kidney failure are also close. This demonstrates that ClinicalBERT captures some clinical semantics.

3.2.3 Quantitative Analysis. We benchmark embedding models using the clinical concept dataset in [22]. The data consists of concept pairs, and the similarity of a pair is rated by physicians, with a score ranging from 1.0 to 4.0 (least similar to most similar). To evaluate representations of clinical text, we calculate the similarity between two concepts' embeddings a and b using cosine similarity,

$$\text{Similarity}(a,b) = \frac{a \cdot b}{\|a\|\|b\|} \tag{3}$$

We calculate the Pearson correlation between physician ratings of medical concept similarity and the cosine similarity between model embeddings. Models with high correlation capture human-rated similarity between clinical terms. Wang et al. [34] conducts a similar evaluation on this data using Word2Vec word embeddings [20] trained on clinical notes, biomedical literature, and Google News. However, this work relies on a private clinical note dataset from The Mayo Clinic to train the Word2Vec model. For a fair comparison with ClinicalBERT, we retrain the Word2Vec model using clinical notes from MIMIC-III. The Word2Vec model is trained on 2.8B words from MIMIC-III with the same hyperparameters as [34]. Word2Vec cannot handle out-of-vocabulary words; we ignore the three medical pairs in the clinical concepts dataset that do not have embeddings (correlation is computed using the remaining 27 medical pairs). Because of this shortcoming, we also train a FastText model [6] on MIMIC-III, which models out-of-vocabulary words using subword units. FastText and Word2Vec are trained on the full MIMIC-III data, so we also pre-train ClinicalBERT on the full data for comparison. Table 2 shows how these models correlate with physician, with ClinicalBERT more accurately correlating with physician judgment.

Table 2: ClinicalBERT captures physician-assessed relationships between clinical terms. The Pearson correlation is computed between the cosine similarity of embeddings learned by models of clinical text and physician ratings of the similarity of medical concepts in the dataset of [22]. These numbers are comparable to the best result, 0.632, from [34].

Model	Pearson correlation
ClinicalBERT	0.670
Word2Vec	0.553
FastText	0.487

3.3 Empirical Study II: 30-Day Hospital Readmission Prediction

The representations learned by ClinicalBERT can help address problems in the clinic. We build a model to predict hospital readmission from clinical notes. Compared to benchmark language models, ClinicalBERT accurately predicts readmission. Further, ClinicalBERT predictions can be interrogated by visualizing attention weights to reveal interpretable patterns in medical data.

3.3.1 Cohort. We select a patient cohort from MIMIC-III using patient covariates. The binary readmit label associated with each patient admission is computed as follows. Admissions where a patient is readmitted within 30 days are labeled $\texttt{readmit} = 1$. All other patient admissions are labeled zero, including patients with appointments within 30 days (to model unexpected readmission). In-hospital death precludes readmission, so admissions with deaths are removed. Newborn patients account for 7,863 admissions. Newborns are in the neonatal intensive care unit, where most undergo testing and are sent back for routine care. This leads to a different distribution of clinical notes and readmission labels; we filter out newborns and focus on non-newborn readmissions. The final cohort contains 34,560 patients with 2,963 positive readmission labels and 42,358 negative labels.

3.3.2 Scalable Readmission Prediction. Patients are often associated with many notes. ClinicalBERT has a fixed length of input sequence, so notes are concatenated and split to this maximum length. Predictions for patients with many notes are computed by binning the predictions on each subsequence. The probability of readmission for a patient is computed as follows. For a patient whose notes are split into n subsequences, ClinicalBERT outputs a probability for each subsequence. The probability of readmission is computed using the predictions for each subsequence:

$$P(\texttt{readmit} = 1 \mid h_{\text{patient}}) = \frac{P^n_{\text{max}} + P^n_{\text{mean}} n/c}{1 + n/c}, \tag{4}$$

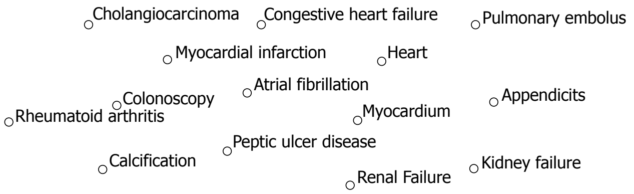

Figure 4: ClinicalBERT reveals interpretable patterns in medical concepts. The model is trained on clinical notes from MIMIC-III, and the embeddings of clinical terms from the dataset in [22] are plotted using the t-distributed stochastic neighbor embedding algorithm for dimensionality reduction [19]. We highlight a subset of the plot centered on a cluster of terms relating to heart conditions such as myocardial infarction, heart failure, and kidney failure.

The scaling factor c controls the influence of the number of subsequences n, and h_{patient}) is the implicit ClinicalBERT representation of all of a patient's notes. The maximum and mean probabilities of readmission over n subsequences are P^n_{max} and P^n_{mean}.

Computing readmission probability using Equation (4) outperforms predictions using the mean for each subsequence by 3–8%. This formula is motivated by observations: some subsequences do not contain information about readmission (such as tokens corresponding to progress reports), whereas others do. The risk of readmission should be computed using subsequences that correlate with readmission, and the effect of unimportant subsequences should be minimized. This is accomplished by using the maximum probability over subsequences. Second, noise in subsequences decreases performance. For example, consider the case where one noisy subsequence has a prediction of 0.8, but all other subsequences have predictions close to zero. Using only the maximum would lead to a false prediction if the maximum is due to noise, so we include the average probability of readmission across subsequences. This leads to a trade-off between the mean and maximum probabilities of readmission in Equation (4). Finally, if there are a large number of subsequences (for a patient with many clinical notes), there is a higher probability of a noisy maximum probability of readmission. This means longer sequences may need a larger weight on the mean prediction. We include this weight as an n/c scaling factor, with c accounting for patients with many notes. The denominator results from normalizing the risk score to the unit interval. The parameter c is selected using the validation set; $c = 2$ was selected.

3.3.3 Evaluation. For validation and testing, the cohort is split into five folds. In each fold 20% is used for validation (10%) and test (10%) sets, with the rest for training. Each model is evaluated using three metrics:

1. Area under the receiver operating characteristic curve (AUROC): the area under the true positive rate versus the false positive rate.
2. Area under the precision-recall curve (AUPRC): the area under the plot of precision versus recall.
3. Recall at precision of 80% (RP80): for readmission prediction, false positives are important. To minimize the number of false positives and hence minimize the risk of alarm fatigue, we fix precision to 80% (or, 20% false positives in the positive class predictions). This threshold is used to calculate recall. This leads to a clinically-relevant metric that enables building models that minimize the false positive rate.

Table 3: ClinicalBERT accurately predicts 30-day readmission using discharge summaries. The mean and standard deviation of 5-fold cross validation is reported. ClinicalBERT outperforms the bag-of-words model, the BI-LSTM, and BERT deep language models.

Model	AUROC	AUPRC	RP80
ClinicalBERT	0.714 ± 0.018	0.701 ± 0.021	0.242 ± 0.111
Bag-of-words	0.684 ± 0.025	0.674 ± 0.027	0.217 ± 0.119
BI-LSTM	0.694 ± 0.025	0.686 ± 0.029	0.223 ± 0.103
BERT	0.692 ± 0.019	0.678 ± 0.016	0.172 ± 0.101

3.3.4 Models. We compare ClinicalBERT to three competitive models. Boag et al. [5] conclude that a bag-of-words model and a long short-term memory (LSTM) model with Word2Vec embeddings work well for predictive tasks on MIMIC-III clinical notes. We also compare to BERT with trainable weights. Training details are in Appendix A.

1. ClinicalBERT: the model parameters include the weights of the encoder network and the learned classifier weights.
2. Bag-of-words: this method uses word counts to represent a note. The 5,000 most frequent words are used as features. Logistic regression with L2 regularization is used to predict readmission.
3. Bidirectional long short-term memory (BI-LSTM) and Word2Vec [27, 14]: a BI-LSTM is used to model words in a sequence. The final hidden layer is used to predict readmission.
4. BERT: this is what ClinicalBERT is based on, but BERT is pretrained not on clinical notes but standard language corpora.

We also compared to ELMo [24], where a standard 1,024-dimensional embedding for each text subsequence is computed and a neural network classifier is used to fit the training readmission labels. The performance was much worse, and we omit these results. This may be because the weights in ELMo are not learned, and the fixed-length embedding may not be able to store the information needed for a classifier to detect signal from long and complex clinical text.

3.3.5 Readmission Prediction with Discharge Summaries. Discharge summaries contain essential information of patient admissions since they are used by the post-hospital care team and by doctors in future visits [32]. The summary may contain information like a patient's discharge condition, procedures, treatments, and significant findings [16]. This means discharge summaries should have predictive value for hospital readmission. Table 3 shows that ClinicalBERT outperforms competitors in terms of precision and recall on a task of readmission prediction using patient discharge summaries.

Table 4: ClinicalBERT outperforms competitive baselines on readmission prediction using clinical notes from early on within patient admissions. In MIMIC-III data, admission and discharge times are available, but clinical notes do not have timestamps. The cutoff time indicates the range of admission durations that are fed to the model from early in a patient's admission. For example, in the 24–48h column, the model may only take as input a patient's notes up to 36h because of that patient's specific admission time. Metrics are reported as the mean and standard deviation of 5 independent runs.

Model	Cutoff time	AUROC	AUPRC	RP80
ClinicalBERT	24–48h	0.674 ± 0.038	0.674 ± 0.039	0.154 ± 0.099
	48–72h	0.672 ± 0.039	0.677 ± 0.036	0.170 ± 0.114
Bag-of-words	24–48h	0.648 ± 0.029	0.650 ± 0.027	0.144 ± 0.094
	48–72h	0.654 ± 0.035	0.657 ± 0.026	0.122 ± 0.106
BI-LSTM	24–48h	0.649 ± 0.044	0.660 ± 0.036	0.143 ± 0.080
	48–72h	0.656 ± 0.035	0.668 ± 0.028	0.150 ± 0.081
BERT	24–48h	0.659 ± 0.034	0.656 ± 0.021	0.141 ± 0.080
	48–72h	0.661 ± 0.028	0.668 ± 0.021	0.167 ± 0.088

3.3.6 Readmission Prediction with Early Clinical Notes. Discharge summaries can be used to predict readmission, but may be written after a patient has left the hospital. Therefore, discharge summaries are not useful for intervention—doctors cannot intervene when a patient has left the hospital. Models that dynamically predict readmission in the early stages of a patient's admission are relevant to clinicians. For the second set of readmission prediction experiments, a maximum of the first 48 or 72 hours of a patient's notes are concatenated. These concatenated notes are used to predict readmission. Since we separate notes into subsequences of the same length, the training set consists of all subsequences up to a cutoff time. The model is tested given notes up to 24–48h or 48–72h of a patient's admission. We do not consider 0-24h cutoff time because there may be too few notes for good predictions. Note that readmission predictions from a model are not actionable if a patient has been discharged. For evaluation, patients that are discharged within the cutoff time are filtered out. Models of readmission prediction are evaluated using the metrics. Table 4 shows that ClinicalBERT outperforms competitors in both experiments. The AUROC and AUPRC results show that ClinicalBERT has more confidence and higher accuracy. At a fixed rate of false alarms, ClinicalBERT recalls more patients that have been readmitted, and its performance increases as the length of admissions increases and the model has access to more clinical notes.

3.3.7 Interpretability. Clinician mistrust of data-driven methods is sensible: predictions from a neural network are difficult to understand for humans, and it is not clear why a model makes a certain prediction or what parts of the data are most informative. ClinicalBERT uses several attention mechanisms which can be used to inspect predictions by visualizing terms correlated with hospital readmission. For a clinical note fed to ClinicalBERT, attention mechanisms compute a distribution over every term in a sentence, given a query term. For a given query vector q computed from an input token, the attention weight distribution is defined as

$$\text{AttentionWeight}(q, K) = \text{softmax}\left(\frac{qK^{\top}}{\sqrt{d}}\right). \qquad (5)$$

The attention weights are used to compute the weighted sum of values. A high attention weight between a query and key token means the interaction between these tokens is predictive of readmission. In the

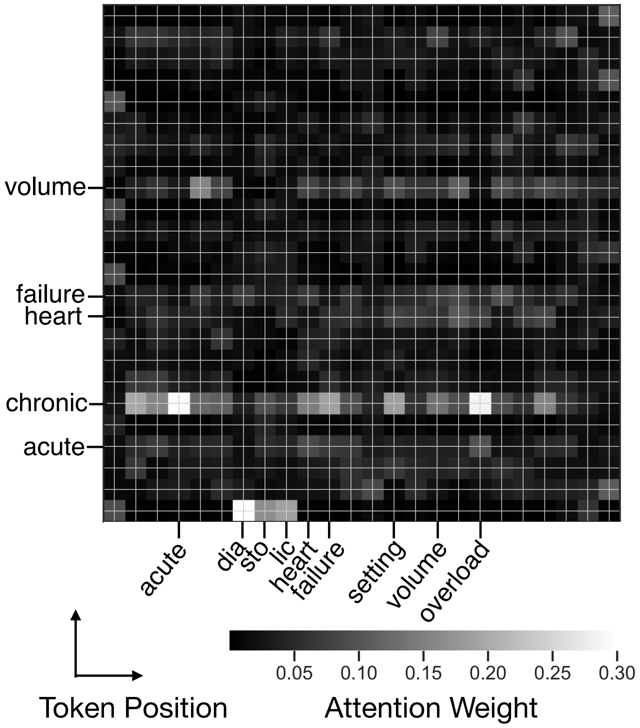

Figure 5: ClinicalBERT provides interpretable predictions, by revealing which terms in clinical notes are predictive of patient readmission. The self-attention mechanisms in ClinicalBERT can be used to interpret model predictions on clinical notes. The input sentence "he has experienced acute chronic diastolic heart failure in the setting of volume overload due to his sepsis." is fed to the model (this sentence is representative of a clinical note found in MIMIC-III). Equation (5) is used to compute a distribution over tokens in this sentence, where every query token is itself a token in the same input sentence. In the panel, we show one of the self-attention mechanisms in ClinicalBERT, and only label terms that have high attention weight. The x-axis labels are query tokens and the y-axis labels are key tokens.

ClinicalBERT encoder, there are 144 self-attention mechanisms (or, 12 multi-head attention mechanisms for each of the 12 transformer encoders). After training, each mechanism specializes to different patterns in clinical notes that are indicative of readmission.

To illustrate, a sentence representative of a MIMIC-III note is fed to ClinicalBERT. Both the queries and keys are the tokens in the sentence. Attention weight distributions for every query are computed using Equation (5) and visualized in Figure 5. The panel shows an attention mechanism that is activated for the word 'chronic' and 'acute' given any query term. This means some attention heads focus on for specific predictive terms, a similar computation to a bag-of-words model. Intuitively, the word 'chronic' is a predictor of readmission.

4 Guidelines on using ClinicalBERT in Practice

ClinicalBERT is pre-trained on MIMIC-III, which consists of patients from ICUs in one Boston hospital. As notes vary by institution and clinical setting (e.g. ICU vs outpatient), to use ClinicalBERT in practice we recommend training ClinicalBERT using the private EHR dataset available at the practitioner's institution. After fitting the model, ClinicalBERT can be used for downstream clinical tasks (e.g. mortality prediction or length-of-stay prediction). We include a tutorial for adapting ClinicalBERT for such downstream classification tasks in the repository.

5 Discussion

We developed ClinicalBERT, a model for learning deep representations of clinical text. Empirically, ClinicalBERT is an accurate language model and captures physician-assessed semantic relationships in clinical text. In a 30-day hospital readmission prediction task, ClinicalBERT outperforms a deep language model and yields a large relative increase on recall at a fixed rate of false alarms. Future work includes engineering to scale ClinicalBERT to capture dependencies in long clinical notes; the max and sum operations in Equation (4) may not capture correlations within long notes. Finally, note that the MIMIC-III dataset we use is small compared to the large volume of clinical notes available internally at hospitals. Rather than using pre-trained MIMIC-III ClinicalBERT embeddings, this suggests that the use of ClinicalBERT in hospitals should entail re-training the model on this larger collection of notes for better performance. The publicly-available ClinicalBERT model parameters can be used to evaluate performance on clinically-relevant prediction tasks based on clinical notes.

6 Acknowledgements

We thank Noémie Elhadad for helpful discussion. Grass icon by Milinda Courey from the Noun Project.

References

[1] E. Alsentzer, J. R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, and M. B. A. McDermott. "Publicly Available Clinical BERT Embeddings". In: *arXiv:1904.03323* (2019).

[2] G. F. Anderson and E. P. Steinberg. "Hospital readmissions in the Medicare population". In: *New England Journal of Medicine* 21 (1984).

[3] D. Banerjee, C. Thompson, C. Kell, R. Shetty, Y. Vetteth, H. Grossman, A. DiBiase, and M. Fowler. "An informatics-based approach to reducing heart failure all-cause readmissions: the Stanford heart failure dashboard". In: *Journal of the American Medical Informatics Association* 3 (2016).

[4] S. Basu Roy, A. Teredesai, K. Zolfaghar, R. Liu, D. Hazel, S. Newman, and A. Marinez. "Dynamic Hierarchical Classification for Patient Risk-of-Readmission". In: *Knowledge Discovery and Data Mining* (2015).

[5] W. Boag, D. Doss, T. Naumann, and P. Szolovits. "What's in a Note? Unpacking Predictive Value in Clinical Note Representations". In: *AMIA Joint Summits on Translational Science* (2018).

[6] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov. "Enriching word vectors with subword information". In: *Transactions of the Association for Computational Linguistics* (2017).

[7] X. Cai, O. Perez-Concha, E. Coiera, F. Martin-Sanchez, R. Day, D. Roffe, and B. Gallego. "Real-time prediction of mortality, readmission, and length of stay using electronic health record data". In: *Journal of the American Medical Informatics Association* 3 (2015).

[8] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad. "Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission". In: *Knowledge Discovery and Data Mining*. 2015.

[9] W. W. Chapman, P. M. Nadkarni, L. Hirschman, L. W. D'Avolio, G. K. Savova, and O. Uzuner. "Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions". In: *Journal of the American Medical Informatics Association* 5 (2011).

[10] B. Chiu, G. Crichton, A. Korhonen, and S. Pyysalo. "How to Train good Word Embeddings for Biomedical NLP". In: *Proceedings of the 15th Workshop on Biomedical Natural Language Processing, ACL 2016* ().

[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *arXiv:1810.04805* (2018).

[12] J. Futoma, J. Morris, and J. Lucas. "A comparison of models for predicting early hospital readmissions". In: *Journal of Biomedical Informatics* (2015).

[13] B. A. Goldstein, A. M. Navar, M. J. Pencina, and J. P. A. Ioannidis. "Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review". In: *Journal of the American Medical Informatics Association* (2017).

[14] S. Hochreiter and J. Schmidhuber. "Long Short-Term Memory". In: *Neural Computation* 8 (1997).

[15] A. E. W. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, and R. G. Mark. "MIMIC-III, a freely accessible critical care database". In: *Scientific Data* (2016).

[16] A. J. Kind and M. A. Smith. "Documentation of mandated discharge summary components in transitions from acute to subacute care". In: *Agency for Healthcare Research and Quality* (2008).

[17] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang. "BioBERT: a pre-trained biomedical language representation model for biomedical text mining". In: *arXiv:1901.08746* (2019).

[18] J. Liu, Z. Zhang, and N. Razavian. "Deep EHR: Chronic Disease Prediction Using Medical Notes". In: *Proceedings of the 3rd Machine Learning for Healthcare Conference*. 2018.

[19] L. van der Maaten and G. Hinton. "Visualizing data using t-SNE". In: *Journal of Machine Learning Research* (2008).

[20] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. "Distributed representations of words and phrases and their compositionality". In: *Advances in Neural Information Processing Systems*. 2013.

[21] C. A. Pedersen, P. J. Schneider, and D. J. Scheckelhoff. "ASHP national survey of pharmacy practice in hospital settings: Prescribing and transcribing—2016". In: *American Journal of Health-System Pharmacy* 17 (2017).

[22] T. Pedersen, S. V. Pakhomov, S. Patwardhan, and C. G. Chute. "Measures of semantic similarity and relatedness in the biomedical domain". In: *Journal of Biomedical Informatics* 3 (2007).

[23] J. Pennington, R. Socher, and C. Manning. "Glove: Global Vectors for Word Representation". In: *EMNLP* (2014).

[24] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. "Deep contextualized word representations". In: *arXiv:1802.05365* (2018).

[25] A. Radford. "Improving Language Understanding by Generative Pre-Training". https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf. 2018.

[26] A. Rajkomar, E. Oren, K. Chen, A. M. Dai, N. Hajaj, M. Hardt, P. J. Liu, X. Liu, J. Marcus, M. Sun, P. Sundberg, H. Yee, K. Zhang, Y. Zhang, G. Flores, G. E. Duggan, J. Irvine, Q. Le, K. Litsch, A. Mossin, J. Tansuwan, D. Wang, J. Wexler, J. Wilson, D. Ludwig, S. L. Volchenboum, K. Chou, M. Pearson, S. Madabushi, N. H. Shah, A. J. Butte, M. D. Howell, C. Cui, G. S. Corrado, and J. Dean. "Scalable and accurate deep learning with electronic health records". In: *NPJ Digital Medicine* 1 (2018).

[27] M. Schuster and K. K. Paliwal. "Bidirectional recurrent neural networks". In: *IEEE Trans. Signal Processing* (1997).

[28] S. Sendelbach and M. Funk. "Alarm fatigue: a patient safety concern". In: *AACN Advanced Critical Care* 4 (2013).

[29] R. Sennrich, B. Haddow, and A. Birch. "Neural Machine Translation of Rare Words with Subword Units". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. 2016.

[30] B. Shickel, P. J. Tighe, A. Bihorac, and P. Rashidi. "Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis". In: *IEEE Journal of Biomedical and Health Informatics* 5 (2018).

[31] Y. Si, J. Wang, H. Xu, and K. Roberts. "Enhancing clinical concept extraction with contextual embeddings". In: *Journal of the American Medical Informatics Association* 11 (2019).

[32] C. Van Walraven, R. Seth, P. C. Austin, and A. Laupacis. "Effect of discharge summary availability during post-discharge visits on hospital readmission". In: *Journal of General Internal Medicine* 3 (2002).

[33] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. "Attention is all you need". In: *Advances in Neural Information Processing Systems*. 2017.

[34] Y. Wang, S. Liu, N. Afzal, M. Rastegar-Mojarad, L. Wang, F. Shen, P. Kingsbury, and H. Liu. "A comparison of word embeddings for the biomedical natural language processing". In: *Journal of Biomedical Informatics* (2018).

[35] W.-H. Weng, K. B. Wagholikar, A. T. McCray, P. Szolovits, and H. C. Chueh. "Medical Subdomain Classification of Clinical Notes Using a Machine Learning-Based Natural Language Processing Approach". In: *BMC Medical Informatics and Decision Making* 1 (2017).

[36] C. Xiao, E. Choi, and J. Sun. "Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review". In: *Journal of the American Medical Informatics Association* 10 (2018).

[37] K.-H. Yu, A. L. Beam, and I. S. Kohane. "Artificial intelligence in healthcare". In: *Nature Biomedical Engineering* 10 (2018).

[38] Y. Zhang, R. Jin, and Z.-H. Zhou. "Understanding bag-of-words model: a statistical framework". In: *International Journal of Machine Learning and Cybernetics* 1 (2010).

[39] Y. Zhang, R. Henao, Z. Gan, Y. Li, and L. Carin. "Multi-Label Learning from Medical Plain Text with Convolutional Residual Models". In: *Proceedings of the 3rd Machine Learning for Healthcare Conference*. 2018.

[40] R. B. Zuckerman, S. H. Sheingold, E. J. Orav, J. Ruhter, and A. M. Epstein. "Readmissions, observation, and the hospital readmissions reduction program". In: *New England Journal of Medicine* 16 (2016).

A Hyperparameters and training details

The parameters are initialized to the BERT Base parameters released by [11]; we follow their recommended hyper-parameter settings. The model dimensionality is 768. We use the Adam optimizer with a learning rate of $2x10^{-}5$. The maximum sequence length supported by the model is set to 512, and the model is first trained using shorter sequences. The details of constructing a sequence are in [11]. For efficient mini-batching that avoids padding mini-batch elements of variable lengths with too many zeros, a corpus is split into multiple sequences of equal lengths. Many sentences are packed into a sequence until the maximum length is reached; a sequence may be composed of many sentences. The next sentence prediction task defined in [11] might more accurately be termed a next sequence prediction task. Our ClinicalBERT model is first trained using a maximum sequence length of 128 for 100,000 iterations on the masked language modeling and next sentence prediction tasks, with a batch size 64. Next, the model is trained on longer sequences of maximum length 512 for an additional 100,000 steps with a batch size of 8. When using text that exists in the test set of the fine-tuning task for pre-training, the training and test set during fine-tuning will not be independent. To avoid this, admissions are split into five folds for independent runs, with four folds for pre-training and training during fine-tuning and the fifth for testing during fine-tuning. Hence, for each independent run, during pre-training, we use all the discharge summaries associated with admissions in the four folds. During fine-tuning for readmission task, ClinicalBERT is trained for three epochs with batch size 56 and learning rate $2x10^-5$. The binary classifier is a three layers neural network of shape 768 x 2048, 2048 x 768, and 768 x 1. We fine-tune ClinicalBERT with three epochs and early stopped on validation loss as the criteria.

For Bi-LSTM, for the input word embedding, the Word2Vec model is used. The Bi-LSTM has 200 output units, with a dropout rate of 0.1. The hidden state is fed into a global max pooling layer and a fully-connected layer with a dimensionality of 50, followed by a rectifier activation function. The rectifier is followed by a fully-connected layer with a single output unit with sigmoid activation function. The binary classification objective function is optimized using the Adam adaptive learning rate (40). The Bi-LSTM is trained for three epochs with a batch size of 64 with early stopping based on the validation loss.

For the empirical study, we use a server with 2 Intel Xeon E5-2670v2 2.5GHZ CPUs, 128GB RAM and 2 NVIDIA Tesla P40 GPUs.

B Preprocessing Notes for Pretraining ClinicalBERT

ClinicalBERT requires minimal preprocessing. First, words are converted to lowercase and line breaks and carriage returns are removed. Then de-identified brackets and remove special characters like ==, – are removed. The next sentence prediction pretraining task described in Section 5 requires two sentences at every iteration. The SpaCy sentence segmentation package is used to segment each note. Since clinical notes don't follow rigid standard language grammar, we find rule-based segmentation has better results than dependency parsing-based segmentation. Various segmentation signs that misguide rule-based segmentators are removed (such as 1.2.) or replaced (M.D., dr. with MD, Dr). Clinical notes can include various lab results and medications that also contain numerous rule-based separators, such as 20mg, p.o., q.d.. To address this, segmentations that have less than 20 words are fused into the previous segmentation so that they are not singled out as different sentences.