arXiv:1904.05342v2 [cs.CL] 11 Apr 2019

ClinicalBert: Modeling Clinical Notes and Predicting Hospital Readmission

Kexin Huang KH2383@NYU.EDU
Courant Institute of Mathematical Sciences
New York University
New York City, NY, USA

Jaan Altosaar ALTOSAAR@PRINCETON.EDU
Princeton University
Princeton, NJ, USA

Rajesh Ranganath RAJESHR@CIMS.NYU.EDU
Courant Institute of Mathematical Sciences, Center for Data Science
New York University
New York City, NY, USA

Abstract

Clinical notes contain information about patients that goes beyond structured data like lab values and medications. However, clinical notes have been underused relative to structured data, because notes are high-dimensional and sparse. This work develops and evaluates representations of clinical notes using bidirectional transformers (ClinicalBert). ClinicalBert uncovers high-quality relationships between medical concepts as judged by humans. ClinicalBert outperforms baselines on 30-day hospital readmission prediction using both discharge summaries and the first few days of notes in the intensive care unit. Code and model parameters are available.[1]

1. Introduction

An electronic health record (EHR) stores patient information; it can save money, time, and lives (Pedersen et al., 2017). Every day, more data gets added to an EHR, so analyses may benefit from machine learning. Machine learning techniques leverage structured features in EHR data, such as lab results and electrocardiography measurements, to uncover patterns and improve predictions (Shickel et al., 2018; Xiao et al., 2018a; Yu et al., 2018). 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 (Boag et al., 2018; Weng et al., 2017; Liu et al., 2018; Afzal et al., 2018). 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

1. Code for reproducing results is available at https://github.com/kexinhuang12345/clinicalBERT and model parameters can be found at http://bit.ly/clinicalbert_weights

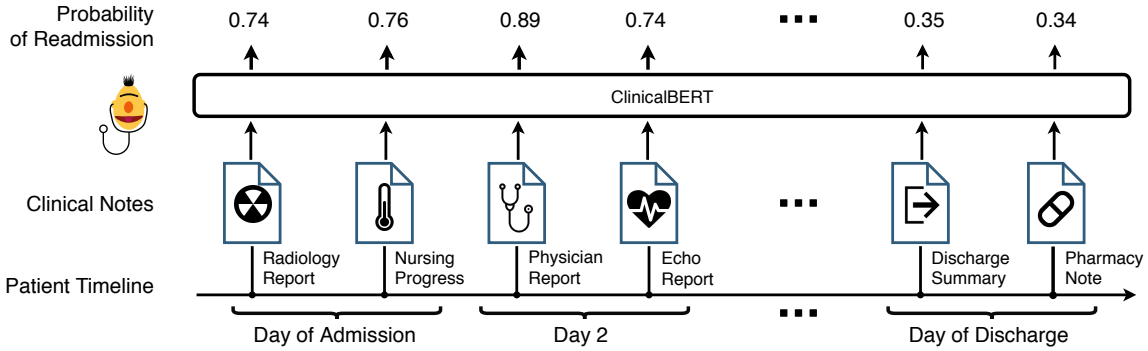

Figure 1: **ClinicalBert learns deep representations of clinical notes that are useful for tasks such as readmission prediction.** In this example, care providers add notes to an electronic health record during a patient's admission, and the model dynamically updates the patient's risk of being readmitted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

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—tools that can make accurate predictions based on clinical notes might be useful in practice.

Hospital readmission lowers patients' quality of life and wastes money (Anderson and Steinberg, 1984; Zuckerman et al., 2016). One estimate puts the financial burden of readmission at 17.9 billion dollars and the fraction of avoidable admissions at 76% (Basu Roy et al., 2015). Accurately predicting readmission has clinical significance both in terms of efficiency and reducing the burden on intensive care unit doctors.

We develop a discharge support model, ClinicalBert, that processes a patient's 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 associated risk score of readmission. This score can help care providers make informed decisions and intervene in advance if needed. ClinicalBert is also readily adapted to other tasks such as diagnosis predictions, mortality risk estimation, or length of stay assessments.

Related Work. Clinical notes use abbreviations, jargon, and have an unusual grammatical structure. Building models that learn useful representations of clinical text is a challenge.

Sager et al. (1995) frame representation learning for clinical notes as machine translation, translating unstructured text to representative sets of words. The bag-of-words model can be used for tasks dependent on individual words (Zhang et al., 2010). Log-bilinear word embedding models such as word2vec have also been used for learning representations of clinical notes (Mikolov et al., 2013; Pennington et al., 2014). Boag et al. (2018) study the performance of the bag-of-words model, word2vec, and a Long Short-Term Memory Network (lstm) model combined with word2vec on various tasks such as diagnosis prediction and mortality risk estimation. Word embedding models such as word2vec are trained us-

ing the local context of individual words, but as clinical notes are long and their words are interdependent (Zhang et al., 2018), these methods cannot capture long-range dependencies.

Natural language processing methods where representations include global, long-range information can yield a boost in performance on various tasks (Peters et al., 2018; Radford, 2018; Devlin et al., 2018). Clinical notes require capturing interactions between distant words. The need to model this long-range structure makes clinical notes suitable for contextual representations like in the bidirectional encoder representations from transformers (BERT) model (Devlin et al., 2018). We develop ClinicalBert by applying BERT to clinical notes. Concurrent to our work, Lee et al. (2019) apply BERT to biomedical literature and Alsentzer et al. (2019) apply BERT on clinical notes and discharge summaries.

Methods to evaluate models of clinical notes are also relevant to ClinicalBert. Wang et al. (2018); Chiu et al. evaluate the quality of biomedical embeddings by computing correlations between doctor-rated relatedness and embedding similarity scores. They also evaluate models through performance on downstream tasks such as information extraction. We adopt similar evaluation techniques in our work.

A good representation of clinical text requires 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. (2015) for comparisons of traditional machine learning methods such as random forests and neural networks on hospital readmission tasks. The majority of work in this area has focused on integrating every possible covariate about a patient into a model. Xiao et al. (2018b) use topic models combined with recurrent neural networks for interpretability and learn clinical concept embeddings for a readmission task. Caruana et al. (2015) develop an interpretable model for readmission prediction based on generalized additive models and highlight the need for intelligible clinical predictions. Rajkomar et al. (2018) predict readmission using Fast Healthcare Interoperability Resources codes from notes, alongside structured information. Most of this previous work uses information at discharge. In this work, we develop a model that can predict readmission dynamically.

Clinical Significance. ClinicalBert shows improved 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 for predicting readmission at any timepoint since a patient was admitted.

To evaluate models on the readmission prediction task, we define a metric motivated by a clinical challenge. Medicine suffers from alarm fatigue (Sendelbach and Funk, 2013). This means useful classification rules for medicine need to have high positive predictive value or precision. We evaluate model performance at a fixed positive predictive value. We show that ClinicalBert has highest recall compared to other popular methods for representing clinical notes.

ClinicalBert can be readily applied to other tasks such as mortality prediction and disease prediction. In addition, self-attention weight output by ClinicalBert can be traced back to understand which elements of clinical notes were relevant to the current prediction. This can be used as an interpretability tool for clinicians.

Technical Significance. We apply the BERT model (Devlin et al., 2018) to clinical notes. Clinical notes are lengthy and numerous, and the computationally-efficient architecture of

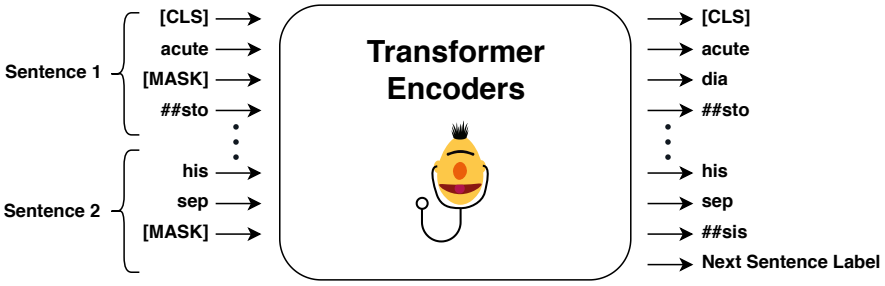

Figure 2: **ClinicalBert learns deep representations of clinical text using two unsupervised language modeling tasks: masked language modeling and next sentence prediction (described in Section 3).** In masked language modeling, a fraction of input tokens are held out for prediction; in next sentence prediction, ClinicalBert predicts whether two input sentences are consecutive.

bert can model long-term dependencies. Compared to a popular model of clinical text, word2vec, 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 a deep language model. We open source ClinicalBert pre-training and readmission model parameters, along with scripts to reproduce results.

2. ClinicalBert

ClinicalBert learns deep representations of clinical text. These deep representations can be used to uncover clinical insights, such as predictions of disease, relationships between treatments and outcomes, or summaries of large volume of texts. ClinicalBert is an application of the bert model (Devlin et al., 2018) to clinical texts; this requires several modifications to address the challenges intrinsic to clinical texts. Specifically, the representations are learned using medical notes and further processed for downstream clinical tasks. As an example, we use the clinical task of hospital readmission prediction.

bert. bert is a deep neural network that uses the transformer encoder architecture (Vaswani et al., 2017) to learn embeddings for text. We omit a detailed description of the architecture; it is described in full in Vaswani et al. (2017). The transformer encoder architecture is based on a self-attention mechanism, and the pre-training objective function for the model is defined using 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.

Clinical Text Embeddings. 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 (Sennrich et al., 2016). In ClinicalBert, a token in a clinical note is computed as the sum of the token embedding, a learned segment embedding, and a position embedding.

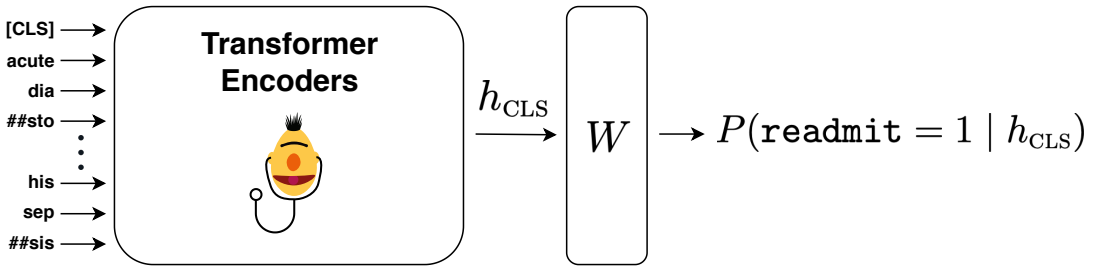

Figure 3: **ClinicalBert models clinical notes and can be readily adapted to clinical tasks such as predicting 30-day readmission.** The model is fed a patient's clinical notes, and the patient's risk of readmission within a 30-day window is predicted using a linear layer applied to the classification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

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, and is used in downstream classification tasks.

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. For a single query, the output of the attention function is a weighted combination of values. The weight of a given value is determined by the interaction of the query and key. Denote a set of queries, keys, and values by Q, K, and V. The attention function is

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

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 in the input sequence (Vaswani et al., 2017). The length and complex patterns in clinical notes make the transformer architecture that uses this self-attention mechanism a good choice. (We describe how this self-attention mechanism leads to interpretability of clinical text in Section 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. However, these two datasets are distinct from clinical notes, where jargon and abbreviations are common and notes have different syntax and grammar than common language in books or encyclopedias. It is hard to understand clinical notes without professional training. ClinicalBert is pre-trained on clinical notes as follows.

ClinicalBert uses the same pre-training tasks as in Devlin et al. (2018). Masked language modeling consists of masking 15% of the input tokens and using the model to predict the

Model	Masked language modeling	Next sentence prediction
ClinicalBert	**86.80%**	**99.25%**
bert	56.80%	80.50%

Table 1: **ClinicalBert improves over bert on two unsupervised language modeling tasks evaluated on a large corpus of clinical text.** We report the accuracy of masked language modeling (predicting held-out tokens) and next sentence prediction (the binary prediction of whether two sentences are consecutive), on the mimic-iii corpus of clinical notes.

masked tokens. In next sentence prediction, two sentences are fed to the model. The model outputs a binary prediction of whether these two sentences are in consecutive order. The pre-training objective function based on the two tasks is the sum of the log-likelihood of the masked tokens and the log-likelihood of the binary variable indicating whether two sentences are consecutive.

Fine-tuning ClinicalBert. After pre-training the model, ClinicalBert is fine-tuned on a task specific to clinical data: readmission prediction. Let `readmit` be a binary indicator of readmission of a patient within the next 30 days. Given clinical notes as input, the output of ClinicalBert is used to predict the probability of readmission:

$$P(\texttt{readmit} = 1 \mid h_{\text{CLS}}) = \sigma(W h_{\text{CLS}})\,, \tag{2}$$

where σ is the sigmoid function, h_{CLS} is the output of the model associated with 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 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 in Section 4, we study its performance in two experiments. First, we find that ClinicalBert outperforms the original bert model in language modeling tasks. In the second experiment, we compare ClinicalBert to popular word embedding models using a clinical word similarity task. The relationships between medical concepts learned by ClinicalBert exhibit better correlation with human evaluation of similarity. Code for training the ClinicalBert model is available at https://github.com/kexinhuang12345/clinicalBERT and model parameter checkpoints are at http://bit.ly/clinicalbert_weights.

Data. We use the public Multiparameter Intelligent Monitoring in Intensive Care (mimic-iii) dataset (Johnson et al., 2016). 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 Appendix B. For pre-training ClinicalBert, we randomly sample 100,000 notes from mimic-iii.

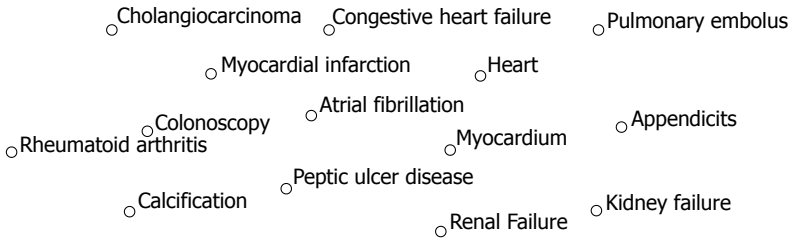

Figure 4: **ClinicalBert reveals interpretable patterns in medical concepts.** The model is trained on clinical notes from MIMIC-III, and the embeddings of clinical terms from the dataset in Pedersen et al. (2017) are plotted using the t-distributed stochastic neighbor embedding algorithm for dimensionality reduction (van der Maaten and Hinton, 2008). The full plot is in Section 5; 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.

ClinicalBert hyperparameters and training. The parameters are initialized to the BERTbase parameters released by Devlin et al. (2018); we follow their recommended hyperparameter settings. The model dimensionality is 768. We use the Adam optimizer (Kingma and Ba, 2015) with a learning rate of 2×10^{-5}. The maximum sequence length supported by the model is set to 512, and the model is first trained using shorter sequences.[2] The model is trained using a maximum sequence length of 128 for 75,000 iterations on the masked language modeling and next sentence prediction tasks, with a batch size 32. Next, the model is trained on longer sequences of maximum length 512 for an additional 75,000 steps with a batch size of 4. The experiments are conducted on Amazon Web Services using a single K80 GPU. The masked language modeling task is illustrated in Figure 7.

Results: Language Modeling. We report the accuracy of the masked language modeling and next sentence prediction tasks on the MIMIC-III data in Table 1. The performance of BERT suffers as the model has not been trained on clinical text. This highlights the need to develop models tailored to clinical data such as ClinicalBert.

Results: Qualitative. We test ClinicalBert on a dataset designed to assess medical term similarity (Pedersen et al., 2017). This dataset consists of 30 pairs of medical terms whose similarity is rated by physicians. Although our model is intended for sequences, we can obtain a feature-based word embedding by feeding the model a sequence that consists of individual tokens corresponding to a medical term. Devlin et al. (2018) conclude that the concatenation or sum of the last four hidden states of the encoders in BERT has the best performance on downstream tasks compared to other combination of hidden states. We use the sum of the last four hidden states of encoders as a representation of medical

2. The details of constructing a sequence are in Devlin et al. (2018). For efficient minibatching that avoids padding minibatch 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 Devlin et al. (2018) might more accurately be termed a next *sequence* prediction task.

Model	Pearson correlation
ClinicalBert	**0.670**
word2vec	0.553
FastText	0.487

Table 2: **ClinicalBert accurately captures relationships between clinical terms as assessed by physicians.** 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 Pedersen et al. (2017) dataset. These numbers are comparable to the best result (0.632) from Wang et al. (2018).

terms. Medical terms are of various lengths in the clinical concept dataset, so the hidden states for each subword unit are summed and divided by the number of subword units. This results in a fixed 768-dimensional vector for each medical term in the dataset. We visualize the similarity of medical terms using dimensionality reduction (van der Maaten and Hinton, 2008). The full plot is in Appendix A; we highlight a cluster of heart-related concepts in Figure 4. ClinicalBert has learned a representation space that groups similar medical concepts. Heart-related concepts such as myocardial infarction, atrial fibrillation, and myocardium are close together; renal failure and kidney failure are also close. ClinicalBert has captured some clinical semantics.

Comparing Representations Using Human Similarity Judgments. We benchmark embedding models using the clinical concept dataset in Pedersen et al. (2017). The dataset consists of concept pairs. The similarity of a pair of concepts 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 $\mathbf{a}$ and $\mathbf{b}$ using cosine similarity,

$$\cos(\theta) = \frac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{a}\|\|\mathbf{b}\|} .$$

We calculate the Pearson correlation between physician ratings of medical concept similarity reported in Pedersen et al. (2017) and the computed cosine similarity between model embeddings. A high correlation score means that a model's embeddings capture human-rated similarity between clinical terms. Wang et al. (2018) conducts a similar evaluation on this data using word2vec word embeddings (Mikolov et al., 2013) trained on clinical notes, biomedical literature, and Google News. However, they use 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 2.8 billion words from mimic-iii with the same hyperparameters as in Wang et al. (2018). word2vec cannot handle out-of-vocabulary words; we ignore the three medical pairs in the clinical concepts dataset that do not have embeddings (the correlation score is computed using the remaining 27 medical pairs). Because of this shortcoming, we also train a FastText model (Bojanowski et al., 2017) on mimic-iii. Fast-

Text can compute embeddings for out-of-vocabulary words using subword embeddings (we use the same hyperparameters as in Wang et al. (2018)).

Results: Clinical Concept Similarity. Table 2 shows the correlations with physician judgments of ClinicalBert and competing models. ClinicalBert more accurately captures the relationships between clinical terms assessed from physician judgments.

4. Empirical Study II: 30-Day Hospital Readmission Prediction

The deep representations learned by ClinicalBert can be used to build clinically-relevant models. We focus on the task of building a model for hospital readmission prediction using clinical notes. Compared to competitive models of language, ClinicalBert accurately predicts readmission. ClinicalBert also reveals interpretable patterns in medical data that can be used to understand its predictions.

Cohort. We select a patient cohort from mimic-iii (the full dataset is described in Section 3) using covariates associated with each patient. For the readmission prediction task, we compute the binary `readmit` label associated with each patient admission as follows. Patient admissions for which the patient is readmitted within 30 days receive a label of `readmit` = 1. All other patient admissions receive a label of zero. This includes patients with scheduled appointments within 30 days (since we are interested in unexpected readmission). We notice that 5,854 admissions are in-hospital deaths. Since death does not imply readmission, we remove these admissions. We also observe that there are 7,863 admissions where the patient is of type `newborn`. These are newborns 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 48,150 negative labels.

4.1. Scalable Readmission Prediction.

Patient are typically associated with many notes. The ClinicalBert model has a fixed maximum length of input sequence. We split notes into subsequences (each subsequence is of the maximum length supported by the model), and define how ClinicalBert makes predictions on long sequences by by binning the predictions on each subsequence. The probability of readmission for a patient is computed as follows. Assume the patient's clinical notes are represented as n subsequences and fed to the model separately; the model outputs a probability for each subsequence. The probability of readmission is computed using the probabilities output for each of these subsequences:

$$P(\texttt{readmit} = 1 \mid h_{\text{patient}}) = \frac{P^n_{\text{max}} + P^n_{\text{mean}} n/c}{1 + n/c}\,, \tag{3}$$

where c is a scaling factor that controls the amount of influence of the number of subsequences n, and h_{patient} is the implicit representation ClinicalBert computes from the entirety of a patient's notes. P^n_{max} is the maximum of probability of readmission across the n subsequences, and P_{mean} is the mean of the probability of readmission across the n subsequences a patient's notes have been split into.

We find that computing readmission probability using Equation (3) consistently outperforms predictions on each subsequence individually by 3–8%. Equation (3) is motivated by observations: some subsequences (such as tokens corresponding to progress reports) do not contain information about readmission, whereas others do. The risk of readmission should be computed using subsequences that correlate with readmission risk, and the effect of unimportant subsequences should be minimized. This is accomplished by using the maximum probability over subsequences. Second, noisy subsequences mislead the model and decrease performance. For example, say there are 4 subsequences with a score lower than 0.3, and one one noisy subsequence with a score of 0.8. Simply using the maximum will not account for cases where the maximum is the noise: this would lead to a false prediction. Therefore, we also include the average probability of readmission across subsequences. This leads to a trade-off between the mean and maximum probabilities of readmission in Equation (3). Finally, if there are a large number of subsequences for a patient with many clinical notes, there is a higher probability of having a noisy maximum probability of readmission. This means longer sequences may need to have a larger weight on the mean prediction. We include this weight as the n/c scaling factor, with c adjusting for patients with many clinical notes. The denominator comes normalizing the final risk score to be in the unit interval. Empirically, we find that $c = 2$ performs best on validation data.

Evaluation. For validation and testing, 10% of the data is held out respectively, and 5-fold cross-validation is conducted. Each model is evaluated using three metrics:

- Area under the receiver operating characteristic curve: the area under the plot of the true positive rate against the false positive rate at various thresholds.
- Area under the precision-recall curve: the area under the plot of precision versus recall at various thresholds.
- Recall at precision of 80%: For the readmission task, false positives are important. To minimize the number of false positives and thus minimize the risk of alarm fatigue, we set the precision to 80% (in other words, 20% false positives out of the predicted positive class) and use the corresponding threshold to calculate recall. This leads to a clinically-relevant metric that enables us to build models that control the false positive rate.

4.2. Models

We compose two baselines based on results from Boag et al. (2018). Boag et al. (2018) conclude that the bag-of-words model and an LSTM model with WORD2VEC embeddings are two strong baselines for predictive tasks on MIMIC-III clinical notes. We compare ClinicalBert with these two methods.

ClinicalBert. The training parameters are the entire encoder network, along with the classifier W. Note that the data labels are imbalanced: negative labels are subsampled to balance the positive `readmit` labels. In every experiment, ClinicalBert is trained for one epoch with batch size 4 and learning rate 2×10^{-5}. The ClinicalBert model settings are the same as in Section 3. The binary classifier is a linear layer of shape 768×1, illustrated in Figure 2.

Model	Area under receiver operating characteristic	Area under precision-recall	Recall at precision of 80%
ClinicalBert	$\mathbf{0.768 \pm 0.027}$	$\mathbf{0.747 \pm 0.029}$	$\mathbf{0.255 \pm 0.113}$
Bag-of-words	0.684 ± 0.025	0.674 ± 0.027	0.217 ± 0.119
biLSTM	0.694 ± 0.025	0.686 ± 0.029	0.223 ± 0.103

Table 3: **ClinicalBert accurately predicts 30-day readmission prediction using discharge summaries.** The mean of 5-fold cross validation is reported along with the standard deviation. ClinicalBert outperforms both the bag-of-words model and the biLSTM deep language model.

Bag-of-words. The bag-of-words model is a simple method that uses word counts to represent a note. We pick the top 5,000 term frequency-inverse document frequency (tf-idf) words as features. This means each note is represented by a 5,000-dimensional vector where each entry is the count of the corresponding vocabulary word occurring in the note. The top 5,000 tf-idf words are computed using the training set. The bag-of-words representation is then computed for every note in the training and test sets. Logistic regression with L2 regularization is used to fit the training readmission labels.

biLSTM and word2vec. Although bag-of-words method is simple and fast, it does not consider the temporal relationship between words in the note. A Bidirectional Long Short-Term Memory Network (biLSTM) (Schuster and Paliwal, 1997; Hochreiter and Schmidhuber, 1997; Graves and Schmidhuber, 2005) is used to build a deep model of relationships between words in a sequence. For the input word embedding, the word2vec model from Section 3 is used. The biLSTM 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 (Kingma and Ba, 2015). The biLSTM is trained for three epochs with a batch size of 64 with early stopping based on the validation loss.

4.3. Readmission Prediction With Discharge Summaries

The discharge summary contains essential information of a patients' stay since it is used by the post-hospital care team and doctors in future visits (Van Walraven et al., 2002). The summary may contain information like patients' discharge conditions, procedures, and treatments, and significant findings (Kind and Smith, 2008). 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 based on patient discharge summaries.

Model	24h–48h	48h–72h
ClinicalBert	**0.673 ± 0.041**	**0.674 ± 0.043**
Bag-of-words	0.648 ± 0.029	0.654 ± 0.035
biLSTM	0.649 ± 0.044	0.656 ± 0.035

(a) Area under receiver operating characteristic

Model	24h–48h	48h–72h
ClinicalBert	**0.670 ± 0.042**	**0.677 ± 0.044**
Bag-of-words	0.650 ± 0.027	0.657 ± 0.026
biLSTM	0.660 ± 0.036	0.668 ± 0.028

(b) Area under precision-recall

Model	24h–48h	48h–72h
ClinicalBert	**0.167 ± 0.090**	**0.171 ± 0.107**
Bag-of-words	0.144 ± 0.094	0.122 ± 0.106
biLSTM	0.143 ± 0.080	0.150 ± 0.081

(c) Recall at precision of 80%

Table 4: **ClinicalBert outperforms competitive baselines on readmission prediction based on clinical notes from early on within patient admissions.** In the mimic-iii data, admission and discharge times are available, but clinical notes do not have timestamps. This is why the table headings show a range; this range shows the cutoff time for notes fed to the model from early on 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 of 5-fold cross validation alongside the standard deviation.

4.4. Readmission Prediction With Early Clinical Notes

Discharge summaries have predictive power for readmission. However, discharge summaries might be written after a patient has left the hospital. Therefore, discharge summaries are not actionable since 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 within a maximum of 72 hours, and the model is tested given only available notes within the first 48 or 72 hours of a patient's admission. Note that readmission predictions from a model are not actionable if a patient has been

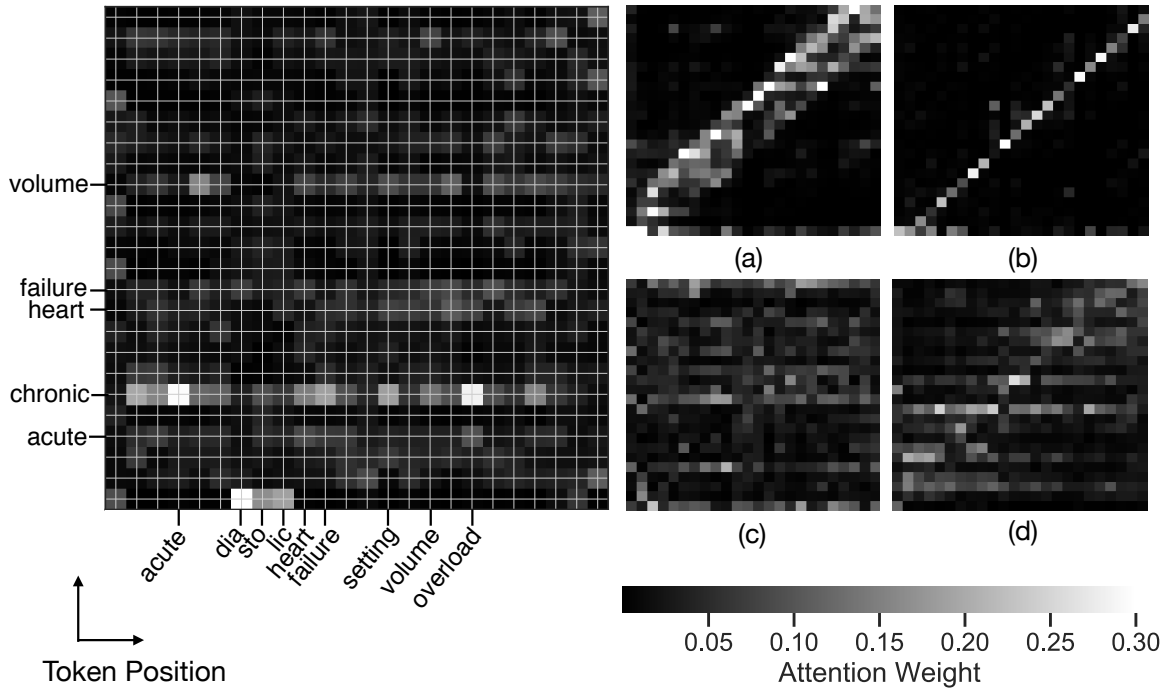

Figure 5: **ClinicalBert provides interpretable predictions, by revealing which terms in clinical notes are predictive of patient readmission.** The self-attention mechanisms in ClinicalBert can be used to interpret model predictions on clinical notes. The input sentence "*he has experienced acute chronic 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 (4) 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 left panel, we show one of the self-attention mechanisms in ClinicalBert, and only label terms that have high attention weight. The x-axis are the query tokens and the y-axis are key tokens. Panels (a), (b), (c), and (d) show other self-attention mechanisms in the model with the same input sentence, showing that they have specialized to various patterns in clinical notes indicative of readmission.

discharged. For testing 48 or 72-hour clinical note readmission prediction, patients that are discharged within 48 or 72 hours (respectively) are filtered out.

The models in Section 4.2 are evaluated using the metrics in Section 4.3. Table 4 shows that ClinicalBert outperforms the baselines in both experiments. The receiver operating characteristic and precision-recall 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. The accuracy of ClinicalBert increases as the length of admissions increases and the model has access to more clinical notes.

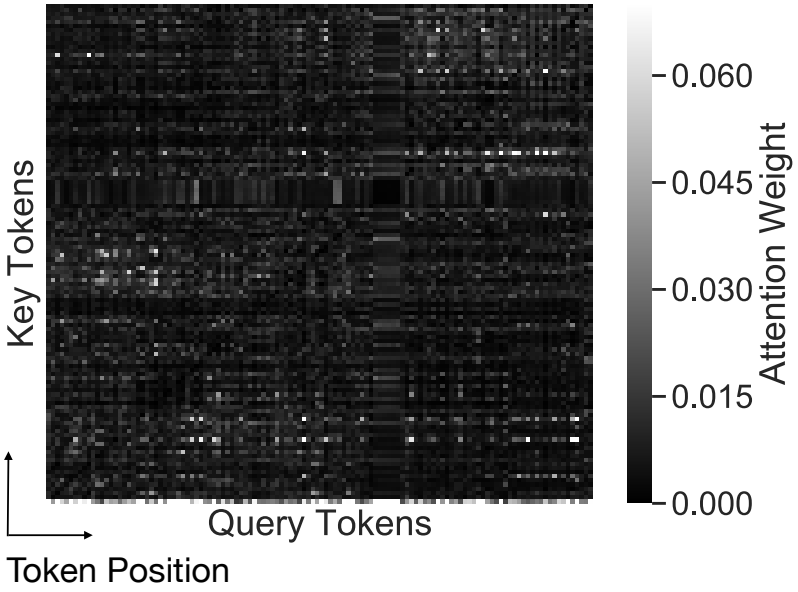

Figure 6: **ClinicalBert captures long-range interactions in clinical notes.** Word position is increasing along both axes. The high attention weights in the upper right and lower left regions of the attention map indicate long-range relationships between clinical terms. The attention mechanism is defined in Section 2

4.5. Interpretable predictions in ClinicalBert

Clinicians may mistrust data-driven methods because of a lack of interpretable predictions: predictions from a neural network are difficult to understand for humans, and it is not clear why a model made a certain prediction or what parts of input data were most informative. ClinicalBert uses several self-attention mechanisms which can be used to inspect its predictions, by visualizing terms correlated with predictions of hospital readmission.

For every clinical note input to ClinicalBert, each self-attention mechanism computes a distribution over every term in a sentence, given a query. For a given query vector q computed from an input token, the attention weight distribution is defined as

$$\text{AttentionWeight}(q, K) = \text{softmax}\left(\frac{qK^\top}{\sqrt{d}}\right). \tag{4}$$

The attention weights are the weights used to compute the weighted sum of values in Equation (1). A high attention weight between a query and key token means the interaction between these tokens is predictive of readmission. In the 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 this, a sentence representative of a MIMIC-III note is input to ClinicalBert. For this sentence, the queries are the tokens in the sentence, and the keys are the tokens in the same sentence. Select attention mechanism distributions for every query are computed

using Equation (4) and visualized in Figure 5. The left panel shows an attention mechanism that is activated for the word *chronic* and *acute* given any query term. It means some heads search for specific predictive terms. This computation is similar to the bag-of-words model. Intuitively, presence of the token associated with the word "chronic" is a predictor of readmission.

The attention mechanism visualized in panel (a) of Figure 5 shows that attention weights that are high for keys within a certain window of a query token. This attention mechanism may focus on local information analogously to the local context window used in word2vec. The attention mechanism in panel (b) shows a pattern shifted below the diagonal: this means that the attention weight is high when the query and key terms are adjacent, which is reminiscent of a bigram model (trigram-type of attention mechanisms have also been observed). The attention weights in (c) are less interpretable. Figure 6 visualizes attention weights for a long clinical note. The large off-diagonal attention weights show that ClinicalBert captures correlation across long ranges of clinical text to make readmission predictions.

5. Discussion

We developed ClinicalBert, a model for learning deep representations of clinical text. Empirically, ClinicalBert is an accurate language model and captures semantic relationships in clinical text as judged by physicians. In a 30-day hospital readmission prediction task, ClinicalBert outperforms a deep language model and yields a 15% relative increase on recall at a fixed rate of false alarms. Future work includes engineering to scale ClinicalBert to capture dependencies in very long clinical notes; the max and sum operations in Equation (3) may not capture correlations within long notes. The publicly-available ClinicalBert model parameters can be used to evaluate performance on clinically-relevant prediction tasks based on clinical notes.

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

References

Naveed Afzal, Vishnu Mallipeddi, Sunghwan Sohn, Hongfang Liu, Rajeev Chaudhry, Christopher G Scott, Iftikhar J Kullo, and Adelaide M Arruda-Olson. Natural language processing of clinical notes for identification of critical limb ischemia. *International Journal of Medical Informatics*, 111:83–89, 2018.

Emily Alsentzer, John R. Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew B. A. McDermott. Publicly available clinical BERT embeddings. *arXiv:1904.03323*, 2019.

Gerard F Anderson and Earl P Steinberg. Hospital readmissions in the medicare population. *New England Journal of Medicine*, 311(21):1349–1353, 1984.

Senjuti Basu Roy, Ankur Teredesai, Kiyana Zolfaghar, Rui Liu, David Hazel, Stacey Newman, and Albert Marinez. Dynamic hierarchical classification for patient risk-of-readmission. *Knowledge Discovery and Data Mining*, pages 1691–1700, 2015.

Willie Boag, Dustin Doss, Tristan Naumann, and Peter Szolovits. What's in a note? Unpacking predictive value in clinical note representations. *AMIA Joint Summits on Translational Science*, 2017:26–34, 05 2018.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017.

Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Knowledge Discovery and Data Mining*, pages 1721–1730, 2015.

Billy Chiu, Gamal Crichton, Anna Korhonen, and Sampo Pyysalo. How to train good word embeddings for biomedical NLP. *Proceedings of the 15th Workshop on Biomedical Natural Language Processing, ACL 2016*, pages 166–174.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv:1810.04805*, 2018.

Joseph Futoma, Jonathan Morris, and Joseph Lucas. A comparison of models for predicting early hospital readmissions. *Journal of Biomedical Informatics*, 56:229 – 238, 2015.

Alex Graves and Jürgen Schmidhuber. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks*, 18(5-6):602–610, 2005.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

Matthew Honnibal and Ines Montani. spaCy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*, 2017.

Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 05 2016.

Amy JH Kind and Maureen A Smith. Documentation of mandated discharge summary components in transitions from acute to subacute care. *Agency for Healthcare Research and Quality*, 2008.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference for Learning Representations*, 2015.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *arXiv:1901.08746*, 2019.

Jingshu Liu, Zachariah Zhang, and Narges Razavian. Deep EHR: Chronic disease prediction using medical notes. In *Proceedings of the 3rd Machine Learning for Healthcare Conference*, pages 440–464, 2018.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119, 2013.

Craig A. Pedersen, Philip J. Schneider, and Douglas J. Scheckelhoff. ASHP national survey of pharmacy practice in hospital settings: Prescribing and transcribing—2016. *American Journal of Health-System Pharmacy*, 74(17):1336–1352, 2017.

Jeffrey Pennington, Richard Socher, and Christoper Manning. Glove: Global vectors for word representation. *EMNLP*, 14:1532–1543, 2014.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *arXiv:1802.05365*, 2018.

Alec 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.

Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M. Dai, Nissan Hajaj, Michaela Hardt, Peter J. Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, Patrik Sundberg, Hector Yee, Kun Zhang, Yi Zhang, Gerardo Flores, Gavin E. Duggan, Jamie Irvine, Quoc Le, Kurt Litsch, Alexander Mossin, Justin Tansuwan, De Wang, James Wexler, Jimbo Wilson, Dana Ludwig, Samuel L. Volchenboum, Katherine Chou, Michael Pearson, Srinivasan Madabushi, Nigam H. Shah, Atul J. Butte, Michael D. Howell, Claire Cui, Greg S. Corrado, and Jeffrey Dean. Scalable and accurate deep learning with electronic health records. *NPJ Digital Medicine*, 1(1):18, 2018.

N Sager, M Lyman, N T Nhan, and L J Tick. Medical language processing: Applications to patient data representation and automatic encoding. *Methods of Information in Medicine*, 34:140–6, 1995.

Mike Schuster and Kuldip K. Paliwal. Bidirectional recurrent neural networks. *IEEE Trans. Signal Processing*, 45:2673–2681, 1997.

Sue Sendelbach and Marjorie Funk. Alarm fatigue: a patient safety concern. *AACN Advanced Critical Care*, 24(4):378–386, 2013.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1715–1725, 2016.

Benjamin Shickel, Patrick James Tighe, Azra Bihorac, and Parisa Rashidi. Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. *IEEE Journal of Biomedical and Health Informatics*, 22(5):1589–1604, 2018.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 11 2008.

Carl Van Walraven, Ratika Seth, Peter C Austin, and Andreas Laupacis. Effect of discharge summary availability during post-discharge visits on hospital readmission. *Journal of General Internal Medicine*, 17(3):186–192, 2002.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

Yanshan Wang, Sijia Liu, Naveed Afzal, Majid Rastegar-Mojarad, Liwei Wang, Feichen Shen, Paul Kingsbury, and Hongfang Liu. A comparison of word embeddings for the biomedical natural language processing. *Journal of Biomedical Informatics*, 87:12 – 20, 2018.

Wei-Hung Weng, Kavishwar B Wagholikar, Alexa T McCray, Peter Szolovits, and Henry C Chueh. Medical subdomain classification of clinical notes using a machine learning-based natural language processing approach. *BMC Medical Informatics and Decision Making*, 17(1):155, 2017.

Cao Xiao, Edward Choi, and Jimeng Sun. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. *Journal of the American Medical Informatics Association*, 25(10):1419–1428, 2018a.

Cao Xiao, Tengfei Ma, Adji B Dieng, David M Blei, and Fei Wang. Readmission prediction via deep contextual embedding of clinical concepts. *PLOS ONE*, 13(4), 2018b.

Kun-Hsing Yu, Andrew L Beam, and Isaac S Kohane. Artificial intelligence in healthcare. *Nature Biomedical Engineering*, 2(10):719, 2018.

Yin Zhang, Rong Jin, and Zhi-Hua Zhou. Understanding bag-of-words model: a statistical framework. *International Journal of Machine Learning and Cybernetics*, 1(1):43–52, 2010.

Yinyuan Zhang, Ricardo Henao, Zhe Gan, Yitong Li, and Lawrence Carin. Multi-label learning from medical plain text with convolutional residual models. In *Proceedings of the 3rd Machine Learning for Healthcare Conference*, pages 280–294, 2018.

Rachael B Zuckerman, Steven H Sheingold, E John Orav, Joel Ruhter, and Arnold M Epstein. Readmissions, observation, and the hospital readmissions reduction program. *New England Journal of Medicine*, 374(16):1543–1551, 2016.

Appendix A. Dimensionality-reduced clinical concept embeddings

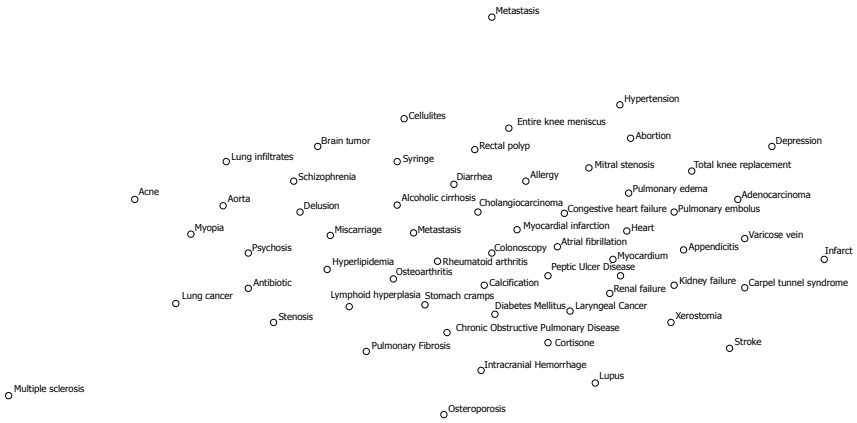

Figure 7: **ClinicalBert captures qualitative relationships among clinical concepts in a database of medical terms from Pedersen et al. (2017).** The medical terms embeddings are computed using the output of ClinicalBert, and visualized using the t-distributed stochastic neighbor embedding algorithm for dimensionality reduction (van der Maaten and Hinton, 2008).

Appendix 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 (Honnibal and Montani, 2017). 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.