An Analysis of Attention over Clinical Notes for Predictive Tasks

Sarthak Jain
Northeastern University
jain.sar@husky.neu.edu

Ramin Mohammadi
Northeastern university
mohammadi.r@husky.neu.edu

Byron C. Wallace
Northeastern University
b.wallace@northeastern.edu

Abstract

The shift to electronic medical records (EMRs) has engendered research into machine learning and natural language technologies to analyze patient records, and to predict from these clinical outcomes of interest. Two observations motivate our aims here. First, unstructured notes contained within EMR often contain key information, and hence should be exploited by models. Second, while strong predictive performance is important, *interpretability* of models is perhaps equally so for applications in this domain. Together, these points suggest that neural models for EMR may benefit from incorporation of *attention* over notes, which one may hope will both yield performance gains and afford transparency in predictions. In this work we perform experiments to explore this question using two EMR corpora and four different predictive tasks, that: (i) inclusion of *attention mechanisms* is critical for neural encoder modules that operate over notes fields in order to yield competitive performance, but, (ii) unfortunately, while these boost predictive performance, it is decidedly less clear whether they provide meaningful support for predictions. Code to reproduce all experiments is available at https://github.com/successar/AttentionExplanation.

1 Introduction

The adoption of electronic medical records (EMRs) has spurred development of machine learning (ML) and natural language processing (NLP) methods that analyze the data these records contain; for a recent survey of such efforts, see (Shickel et al., 2018). Key information for downstream predictive tasks (e.g., forecasting whether a patient will need to be readmitted within 30 days) may be contained within unstructured notes fields (Boag et al., 2018; Jin et al., 2018).

In this work we focus on the modules within neural network architectures responsible for encoding text (notes) into a fixed-size representation for consumption by downstream layers. Patient histories are often long and may contain information mostly irrelevant to a given target. Encoding this may thus be difficult, and text encoder modules may benefit from *attention mechanisms* (Bahdanau et al., 2014), which may be imposed to emphasize relevant tokens.

In addition to mitigating noise introduced by irrelevant tokens, attention mechanisms are often seen as providing interpretability, or insight into model behavior. However, recent work (Jain and Wallace, 2019) has argued that treating attention as explanation may, at least in some cases, be misguided. Interpretability is especially important for clinical tasks, but incorrect or misleading rationales supporting predictions may be particularly harmful in this domain; this motivates our focused study in this space.

To summarize, our **contributions** are as follows. First, we empirically investigate whether incorporating standard attention mechanisms into RNN-based text encoders improves the performance of predictive models learned over EMR. We find that they do; inclusion of standard additive attention mechanism in LSTMs consistently yields absolute gains of $\sim$10 points in AUC, compared to an LSTM without attention.[1] Second, we evaluate the induced attention distributions with respect to their ability to 'explain' model predictions. We find mixed results here, similar to (Jain and Wallace, 2019): attention distributions correlate only weakly (though almost always significantly) with

[1] Indeed, across both corpora and all tasks considered, inattentive LSTMs perform considerably worse than logistic regression and bag-of-words (BoW); introducing attention makes the neural variants competitive, but not decisively better. We hope to explore this point further in future work.

Proceedings of the 2nd Clinical Natural Language Processing Workshop, pages 15–21
Minneapolis, Minnesota, June 7, 2019. ©2019 Association for Computational Linguistics

gradient measures of feature importance, and we are often able to identify very different attention distributions that nonetheless yield equivalent predictions. Thus, one should not in general treat attention weights as meaningful explanation of predictions made using clinical notes.

2 Models

We experiment with multiple standard encoding architectures, including: (i) a standard BiLSTM model; (ii) a convolutional model, and (iii) an embedding projection based model. We couple each of these with an attention layer, following (Jain and Wallace, 2019). Concretely, each encoder yields hidden state vectors $\{h_1, ..., h_T\}$, and an attention distribution $\{\alpha_1, ..., \alpha_T\}$ is induced over these according to a scoring function ϕ: $\hat{\boldsymbol{\alpha}} = \text{softmax}(\phi(\mathbf{h})) \in \mathbb{R}^T$. In this work we consider *Additive* similarity functions $\phi(\mathbf{h}) = \mathbf{v}^T \tanh(\mathbf{W_1 h} + \mathbf{b})$ (Bahdanau et al., 2014), where $\mathbf{v}, \mathbf{W_1}, \mathbf{b}$ are model parameters. Predictions are made on the basis of induced representations: $\hat{y} = \sigma(\boldsymbol{\theta} \cdot h_\alpha) \in \mathbb{R}^{|\mathcal{Y}|}$, where $h_\alpha = \sum_{t=1}^{T} \hat{\alpha}_t \cdot h_t$ and $\boldsymbol{\theta}$ are top-level discriminative (e.g., softmax) parameters.

3 Datasets and Tasks

We consider five tasks over two independent EMR datasets. The first EMR corpus is MIMIC-III (Johnson et al., 2016), a publicly available set of records from patients in the Intensive Care Unit (ICU). We follow prior work in modeling aims and setup on this dataset. Specifically we consider the following predictive tasks on MIMIC.

1. **Readmission**. The task here is to predict patient readmission within 30 days of discharge or transfer from the ICU. We follow the cohort selection of (Lin et al., 2018). We assume the model has access to all notes from patient admission up until the discharge or transfer from the ICU (the point of prediction).

2. **Retrospective 1-yr mortality**. We aim to predict patient mortality within one year. In this we follow the experimental setup of (Ghassemi et al., 2014). The model is provided all notes up until patient discharge (excluding the discharge summary).

3. **Phenotyping**. Here we aim to predict the top 25 acute care phenotypes for patients (associated at discharge with the admission). For this we again rely on the framing established in prior work (Harutyunyan et al., 2017). The model has access to all notes from admission up until the end of the ICU stay. Note that this may be viewed as a multilabel classification task, similar to (Harutyunyan et al., 2017; Lipton et al., 2015).

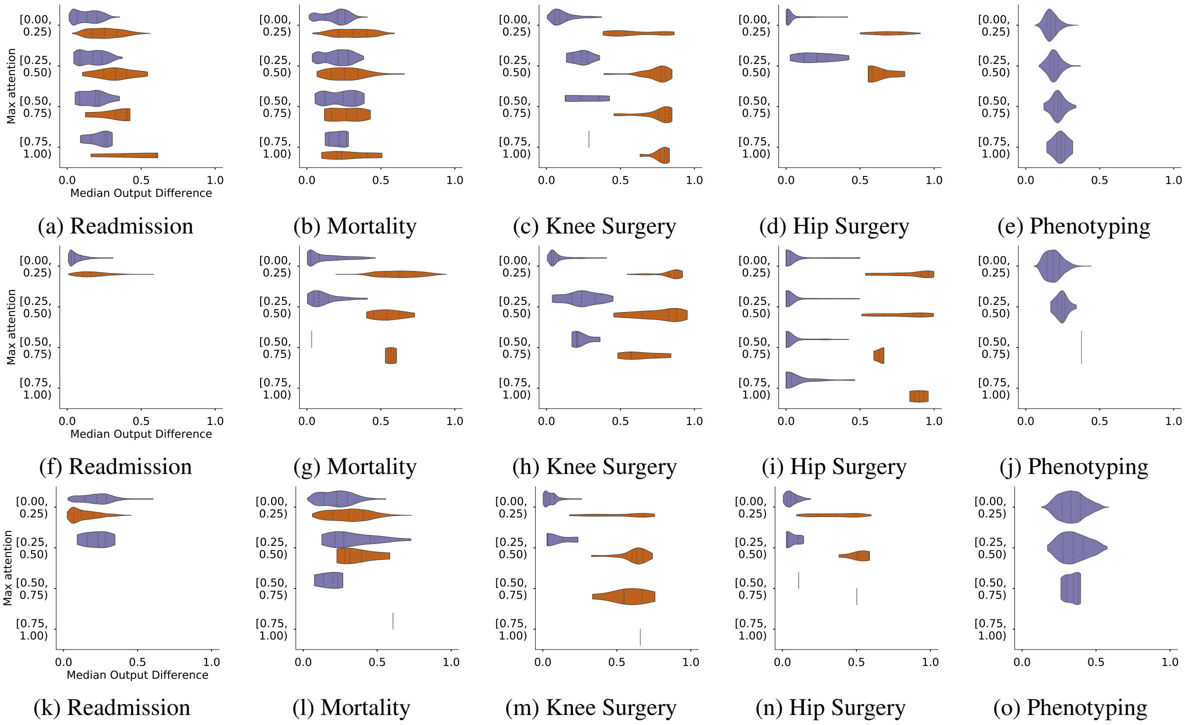
The second EMR dataset we use comprises records for 7174 patients from Mass General Hospital who underwent hip or knee arthroplasty procedures. Use of this data was approved by an Institutional Review Board (IRB protocol number 2016P002062) at Partners Healthcare.

1. **Predicting Hip and Knee Surgery Complications**. We consider patients who underwent hip or knee arthroplasty procedure; we aim to classify these patients with respect to whether or not they will be readmitted within 30 days due to surgery-related complications. We run experiments over hip and knee surgery patients separately.

4 Experiments

Following the analysis of (Jain and Wallace, 2019) but focusing on clinical tasks, we perform a set of experiments on these corpora that aim to assess the degree to which attention mechanisms aid (or hamper) predictive performance, and the degree to which the induced attention weights might be viewed as providing explanations for predictions.

The latter can be assessed in many ways, depending on one's view of interpretability. To address the question of whether it is reasonable to treat attention as providing interpretability broadly, we perform experiments that interrogate multiple properties we might expect these weights to exhibit if so. Specifically, we: probe the degree to which attention weights correlate with alternative gradient-based feature importance measures, which have a more straight-forward interpretation (Ross et al., 2017; Li et al., 2016); evaluate whether we are able to identify 'counterfactual' attention distributions that change the attention weights (focus) but not the prediction; and, in an exercise novel to the present work, we consider replacing attention weights with log odds scores from a logistic regression (linear) model. We provide a web interface to interactively browse the plots for all datasets, model variants, and experiment types: `https://successar.github.io/AttentionExplanation/docs/`.

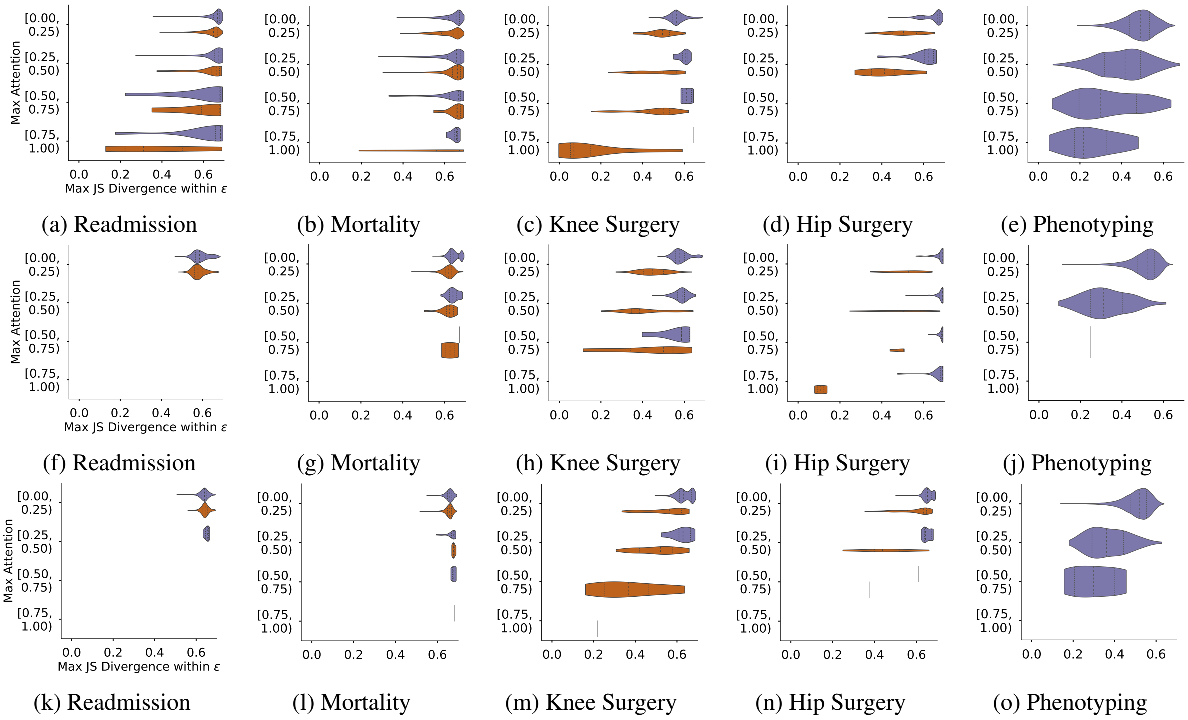


(a) Readmission (b) Mortality (c) Knee Surgery (d) Hip Surgery (e) Phenotyping

(f) Readmission (g) Mortality (h) Knee Surgery (i) Hip Surgery (j) Phenotyping

(k) Readmission (l) Mortality (m) Knee Surgery (n) Hip Surgery (o) Phenotyping

Figure 1: **Median change in output $\Delta\hat{y}^{med}$** (x) densities in relation to the **max attention** (**$\max\hat{\alpha}$**) (y) obtained by randomly permuting instance attention weights. Colors denote classes: negative (🟪) and positive (🟧); phenotyping (e) is not binary. **Top row shows results for BiLSTM encoders; middle for CNNs; bottom for Embedding Projection.**



(a) Readmission (b) Mortality (c) Knee Surgery (d) Hip Surgery (e) Phenotyping

(f) Readmission (g) Mortality (h) Knee Surgery (i) Hip Surgery (j) Phenotyping

(k) Readmission (l) Mortality (m) Knee Surgery (n) Hip Surgery (o) Phenotyping

Figure 2: Densities of **maximum JS divergences (ϵ-max JSD)** (x-axis) as a function of the **max attention** (y-axis) in each instance for obtained between original and adversarial attention weights. Colors are as above. **Top row shows results for BiLSTM encoders; middle for CNNs; bottom for Embedding Projection.**

Dataset	Class	Mean ± Std.	Sig. Frac.
LSTM Encoder			
Readmission	0	0.37 ± 0.04	1.00
	1	0.38 ± 0.05	1.00
Mortality	0	0.33 ± 0.05	1.00
	1	0.35 ± 0.06	1.00
Knee Surgery	0	0.38 ± 0.07	1.00
	1	0.49 ± 0.08	1.00
Hip Surgery	0	0.24 ± 0.07	1.00
	1	0.33 ± 0.09	1.00
Phenotyping	Overall	0.24 ± 0.06	1.00
Projection Encoder			
Readmission	0	0.65 ± 0.03	1.00
	1	0.64 ± 0.03	1.00
Mortality	0	0.76 ± 0.02	1.00
	1	0.76 ± 0.02	1.00
Knee Surgery	0	0.65 ± 0.05	1.00
	1	0.60 ± 0.06	1.00
Hip Surgery	0	0.59 ± 0.09	1.00
	1	0.55 ± 0.09	1.00
Phenotyping	Overall	0.89 ± 0.02	1.00

Table 1: Mean and std. dev. of correlations between gradient importance measures and attention weights. *Sig. Frac.* columns report the fraction of instances for which this correlation is statistically significant.

4.1 Gradient Experiments

To evaluate correlations between attention weights and gradient based feature importance scores, we compute Kendall-τ measure (Table 1) between attention scores and gradients with respect to the tokens comprising documents. Across both corpora and all tasks we observe only a modest correlation between the two for BiLSTM model (the projection based model have higher correspondence, which is expected for such simple architectures). This may be problematic for attention as an explanatory mechanism, given the explicit relationship between gradients and model outputs. (Although we note that gradient based methods themselves pose difficulty with respect to interpretation (Feng et al., 2018)).

4.2 Counterfactual Experiments

We investigate if model predictions *would* have differed, had the model attended to different words (i.e., under *counterfactual* attention distributions).

We follow the two strategies from (Jain and Wallace, 2019) for constructing counterfactual attention distributions. In the first we randomly permute the empirical weights obtained from the attention module prior to inducing the weighted representation $\mathbf{h}_\alpha$. We repeat this process 100 times and record the median change in output.

The second strategy is *adversarial*; we explicitly aim to identify attention weights that are maximally different from the observed weights, with the constraint that this does not change the model output by more some small value ϵ. In both cases, all other model parameters are held constant.

Model	ROC AUC	PR AUC
Readmission		
LR + BoW	0.70	0.29
LSTM	0.63	0.22
LSTM + Additive Attention	0.71	0.30
LSTM + Additive Attention (Log Odds at Test)	0.69	0.26
LSTM + Log Odds Attention	0.71	0.29
Mortality		
LR + BoW	0.82	0.46
LSTM	0.74	0.29
LSTM + Additive Attention	0.83	0.47
LSTM + Additive Attention (Log Odds at Test)	0.80	0.41
LSTM + Log Odds Attention	0.82	0.42
Knee Surgery Complication		
LR + BoW	0.80	0.39
LSTM	0.66	0.18
LSTM + Additive Attention	0.79	0.35
LSTM + Additive Attention (Log Odds at Test)	0.81	0.34
LSTM + Log Odds Attention	0.81	0.38
Hip Surgery Complication		
LR + BoW	0.76	0.32
LSTM	0.63	0.16
LSTM + Additive Attention	0.75	0.24
LSTM + Additive Attention (Log Odds at Test)	0.74	0.26
LSTM + Log Odds Attention	0.78	0.29
Phenotyping		
LR + BoW	0.86	0.59
LSTM	0.78	0.41
LSTM + Additive Attention	0.86	0.58
LSTM + Additive Attention (Log Odds at Test)	0.81	0.48
LSTM + Log Odds Attention	0.85	0.56

Table 2: Predictive results across all datasets and tasks using different models and attention variants.

In Figures 1 and 2, we observe that predictions are unchanged under alternative attention configurations in a significant majority of cases across all architectures. Thus, attention cannot be viewed casually in the sense of 'the model made these predictions *because* these words were attended to'. Alternative attention distributions that yield equivalent predictions would seem to be equally plausible under the view of attention as explanation.

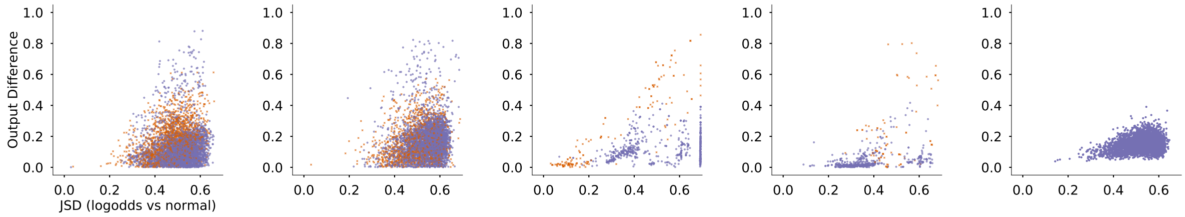
4.3 Log Odds Experiments

As a novel exercise, we also consider swapping log-odds scores for features (from an LR model operating over BoW) in for attention weights in BiLSTM model. Specifically, we induce a 'log odds attention' over an input by substituting the absolute value of log odds (as estimated via LR) of

Original vs Adversarial Attention Difference : Sed dolorem sed adipisci ipsum dolor dolorem. Ut adipisci magnam tempora. Modi # eius : tempora change ipsum adipisci tempora tracheobronchomalacia quaerat dolor. Numquam est dolore labore est neque. respiratory failure Ipsum quiquia etincidunt labore modi. Dolorem aliquam dolore amet. Amet est consectetur modi neque. Porro respiratory failure etincidunt quaerat est neque dolor quaerat. Est quaerat est adipisci ipsum. Sit dolore quisquam ipsum non neque quiquia aliquam. Ut ipsum adipisci labore tempora quaerat tempora labore. Ipsum numquam voluptatem consectetur. Aliquam voluptatem , eius numquam. Velit generalized ut non numquam magnam sed modi. Consectetur porro . heart etincidunt eius consectetur , quaerat amet. Amet dolorem is difficult dolor consectetur etincidunt sed effusions quiquia aliquam. Porro etincidunt dolore labore no dolore dolorem aliquam. Tempora etincidunt quisquam aliquam numquam eius ut. tracheostomy Modi modi amet voluptatem
Original Output: 0.694 **Adversarial Output:** 0.699

Original vs Log Odds Attention Difference : Non magnam quiquia magnam magnam quaerat. Ut etincidunt magnam voluptatem velit eius. Dolorem dolorem velit dolor porro ut etincidunt. Consectetur dolor voluptatem cystic brain mass quaerat surgical resection est magnam etincidunt. Ipsum neque dolorem sed consectetur est. Magnam modi voluptatem dolorem tempora sed ut. Dolore dolor tempora eius aliquam quisquam. Dolor quisquam eius sed labore dolore sit velit. Magnam aliquam quisquam numquam. Aliquam sed sed modi neque. Dolor chronic quiquia voluptatem adipisci quaerat adipisci. . . . . . Magnam velit quaerat adipisci. Ut cystic brain mass adipisci velit modi. Sed aliquam astrocytoma est porro. Labore resection eius voluptatem sit quisquam consectetur modi. Est ipsum tumor dolore
Original Output: 0.798 **Log Odds Output :** 0.800

Figure 3: Heatmaps showing difference in Original and counterfactual attention distributions over clinical notes from MIMIC, where we have replaced text with *lorem ipsum* for all but the most relevant tokens in order to preserve privacy (red implies counterfactual attention is higher and blue vice-versa). These show different cases where we can significantly change the attention distribution (either **adversarial (Top)** or using **Log Odds (Bottom)** while barely affecting the prediction.



(a) Readmission (b) Mortality (c) Knee Surgery (d) Hip Surgery (e) Phenotyping

Figure 4: Change in output (y-axis) using original attention vs Log Odds attention during predictions against JSD between these two distributions (x-axis). These results are for LSTM encoders.

the word present at each position and passing this through a softmax: $\boldsymbol{\alpha}^{LO} = \text{softmax}_t(\{\beta_{w_t}\}_{t=1}^{T})$ where w_t is the word at position t and β are log-odds estimates.

These scores enjoy a clear interpretation under a linear regime. We thus explore two ways of using them with attentive neural models: (1) Swapping in these in as attention weights place of $\mathbf{h}_\alpha$ at test (prediction) time; (2) Use the (fixed) 'log-odds attention' during training, in place of learning the attention distribution end-to-end.

Table 2 shows that using log odds attention at test time does not degrade the performance significantly in most datasets (and actually improves performance for the Knee Surgery Complications task). Similarly, using log odds attention during training also yields similar performance to standard attention variants. But as we see in Figure 4, log odds attention distributions can differ considerably from learned attention distributions, again highlighting the difficulty of interpreting attention weights.

5 Discussion and Conclusions

Across two EMR datasets and five predictive tasks, we have shown that (i) attention mechanisms substantially boost the performance of LSTM text encoders passed over clinical notes, but, (ii) treating attention weights as 'explanations' for predictions is unwarranted. The latter confirms that the recent general findings of (Jain and Wallace, 2019) hold in the clinical domain; this is important because interpretability in this space is critical for obvious reasons.

We hope that this paper inspires work on transparent attention mechanisms for models that make predictions on the basis of EMR.

Acknowledgments

This work was supported by the Army Research Office (ARO), award W911NF1810328.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Willie Boag, Dustin Doss, Tristan Naumann, and Peter Szolovits. 2018. What‘s in a note? unpacking predictive value in clinical note representations. *AMIA Summits on Translational Science Proceedings*, 2017:26.

Shi Feng, Eric Wallace, Alvin Grissom II, Pedro Rodriguez, Mohit Iyyer, and Jordan Boyd-Graber. 2018. Pathologies of neural models make interpretation difficult. In *Empirical Methods in Natural Language Processing*.

Marzyeh Ghassemi, Tristan Naumann, Finale Doshi-Velez, Nicole Brimmer, Rohit Joshi, Anna Rumshisky, and Peter Szolovits. 2014. Unfolding physiological state: Mortality modelling in intensive care units. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 75–84. ACM.

Hrayr Harutyunyan, Hrant Khachatrian, David C Kale, and Aram Galstyan. 2017. Multitask learning and benchmarking with clinical time series data. *arXiv preprint arXiv:1703.07771*.

Sarthak Jain and Byron C Wallace. 2019. Attention is not explanation. *arXiv preprint arXiv:1902.10186*.

Mengqi Jin, Mohammad Taha Bahadori, Aaron Colak, Parminder Bhatia, Busra Celikkaya, Ram Bhakta, Selvan Senthivel, Mohammed Khalilia, Daniel Navarro, Borui Zhang, et al. 2018. Improving hospital mortality prediction with medical named entities and multimodal learning. *arXiv preprint arXiv:1811.12276*.

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3:160035.

Jiwei Li, Will Monroe, and Dan Jurafsky. 2016. Understanding neural networks through representation erasure. *arXiv preprint arXiv:1612.08220*.

Yu-Wei Lin, Yuqian Zhou, Faraz Faghri, Michael J Shaw, and Roy H Campbell. 2018. Analysis and prediction of unplanned intensive care unit readmission using recurrent neural networks with long short-term memory. *bioRxiv*.

Zachary C Lipton, David C Kale, Charles Elkan, and Randall Wetzel. 2015. Learning to diagnose with lstm recurrent neural networks. *arXiv preprint arXiv:1511.03677*.

Sampo Pyysalo, Filip Ginter, Hans Moen, Tapio Salakoski, and Sophia Ananiadou. 2013. Distributional semantics resources for biomedical text processing. *Proceedings of LBM*, pages 39–44.

Andrew Slavin Ross, Michael C Hughes, and Finale Doshi-Velez. 2017. Right for the right reasons: Training differentiable models by constraining their explanations. *arXiv preprint arXiv:1703.03717*.

Benjamin Shickel, Patrick James Tighe, Azra Bihorac, and Parisa Rashidi. 2018. 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.

An Analysis of Attention over Clinical Notes for Predictive Tasks: Appendix

A Dataset Statistics

Task	$\|V\|$	*Avg. length*	*Train size*	*Test size*
Readmission	36464	3865	23790 / 5499	4265 / 735
Mortality	34030	3901	21347 / 4675	4323 / 677
Hip Surgery Complications	10842	2624	3281 / 369	719 / 75
Knee Surgery Complications	10842	2586	2664 / 324	582 / 48
Phenotyping	10842	3641	31075	5000

Table 3: Dataset characteristics. For train and test size, we list the cardinality for each class, where applicable: 0/1 for binary classification and overall for multilabel. Average length is in tokens.

The Phenotypes studied in Phenotyping task are -

Acute and unspecified renal failure, Acute cerebrovascular disease, Acute myocardial infarction, Cardiac dysrhythmias, Chronic kidney disease, Chronic obstructive pulmonary disease and bronchiectasis, Complications of surgical procedures or medical care, Conduction disorders, Congestive heart failure - nonhypertensive, Coronary atherosclerosis and other heart disease, Diabetes mellitus with complications, Diabetes mellitus without complication, Disorders of lipid metabolism, Essential hypertension, Fluid and electrolyte disorders, Gastrointestinal hemorrhage, Hypertension with complications and secondary hypertension, Other liver diseases, Other lower respiratory disease, Other upper respiratory disease, Pleurisy - pneumothorax - pulmonary collapse, Pneumonia (except that caused by tuberculosis or sexually transmitted disease), Respiratory failure - insufficiency - arrest (adult), Septicemia (except in labor), Shock .

B Model Details

For all datasets, we use spaCy for tokenization. We map out of vocabulary words to a special `<unk>` token and map any word with numeric characters to 'qqq'. Each word in the vocabulary was initialized using pretrained embeddings (Pyysalo et al., 2013). We initialize words not present in the vocabulary using samples from a standard Gaussian ($\mu = 0$, $\sigma^2 = 1$).

B.1 BiLSTM

We use an embedding size of 300 and hidden size of 128 for all datasets. The model was regularized with L_2 regularization ($\lambda = 10^{-5}$) applied to all parameters. We use a sigmoid activation function for all binary classification tasks. We treat each phenotype classification as binary classification and take the mean loss over labels during training. We trained the model using maximum likelihood loss function with Adam Optimizer with default parameters in PyTorch.

B.2 CNN

We use an embedding size of 300 and 4 kernels of sizes [1, 3, 5, 7], each with 64 filters, giving a final hidden size of 256. We use ReLU activation function on the output of the filters. All other configurations remain same as BiLSTM.

B.3 Average

We use the embedding size of 300 and a projection size of 256 with ReLU activation on the output of the projection matrix. All other configurations remain same as BiLSTM.