Predicting 30-day Hospital Readmission Using MIMIC-III Clinical Notes

Ryan Belfer
ryan.belfer@berkeley.edu

Nicholas Lovejoy
nflovejoy@berkeley.edu

Abstract

Nearly one in five Medicare patients are readmitted to the hospital after they being discharged due to complications experienced at home. Not only is this an indicator of low quality care, but health insurers levy significant penalties on hospitals with too many unplanned readmissions. One of the key problems with hospital teams trying to reduce readmission rates is the lack of early warning signs available in structured data contained in the patient's electronic medical record (Hosseinzadeh et al., 2013), (Sushmita et al., 2016). Much of the valuable information regarding a patient's course of treatment and the severity of their ailment is contained in large quantities of unstructured, free-text clinical notes. As such, the team sought to use these clinical notes as a basis for predicting readmission probability at various time intervals after a patient's initial admission to the hospital. This work represents a novel approach to preparing the data in a way that captures the maximum amount of information, and offers best-in-class performance using a fine-tuned Bio-Clinical BERT model to obtain an AUROC score of 0.693. An attempt was made to use the logit outputs as inputs to separate neural networks, but a simple evaluation formula was able to outperform these networks by about 5%.

1 Introduction

Hospitals have historically run off a "fee for service" business model, where they are reimbursed based on the quantity of service provided. This financial model runs the risk of incentivising high volume, rather than high value healthcare. To combat this potentially harmful inventive structure, health insurers are increasingly moving towards what's known as a "value based reimbursement" model where reimbursement is made based on the quality of the care provided. One such program is the Hospital Readmission Reduction Program (HRRP) run by the Centers for Medicare and Medicaid Services (CMS). This program penalizes hospitals that have unplanned readmissions of patients within 30 days of discharge by reducing the amount of reimbursement by up to 3%. A readmission can happen when a patient is discharged from the hospital, but then subsequently becomes ill or does not heal as expected. The motivation behind this program is to ensure that patients are being discharged at the appropriate time, and as such, reduce patient mortality and improve patient outcomes.

This HRRP program, and others like it, have driven hospital administrators to try to identify patients who may be at risk for complications once discharged. One such model is the commonly use LACE Index, which uses data regarding the length of stay, acuity of admission, comorbidity of the patient and emergency department utilization within the last six months. More recently, researchers have used machine learning techniques to extract predictions from structured data contained in a patient's medical record . Some researchers have also applied more novel Deep Learning approaches to extracting accurate readmission predictions from structured EMR data such as contextual embeddings (Xiao et al.) and convolutional neural networks (Wang et al., 2018).

While many of these efforts are highly performant, none of them make use of the vast wealth of information contained in clinical notes. Because of discrepancies in how individual hospitals - and even individual caregivers - write clinical notes, there is significant heterogeneity in the form and structure of clinical notes. This, combined with the unique ontology and abbreviations (Figure 1) of clinical notes has proved to be a barrier to using these aspects of a patient's medical record as a proving ground for predictive modeling.

Early identification of readmission risk is critical

> Percutaneous coronary intervention, in [year] anatomy as follows: Patent SVG to OM1, patent SVG to PDA which filled the distal PDA as well as the R-PL via a jump segment. Stump occlusion of a graft presumably to the right system as well as one stump that could be documented of a graft to the left. Other SVG's were not able to be selectively engaged. Supravalvular aortography demonstrated no other patent grafts. Patent LIMA to mid-LAD, which also back-perfused the diagonal via a patent jump graft that was interposed between the LAD and the diagonal.

Figure 1: Example section of a clinical note from the Intensive Care Unit (ICU)

to modifying a patient's care plan while they are still in the hospital and ensuring that they receive the care they need before being discharged home. For this reason, tools which allow caregivers to understand readmission risk as soon as possible are most valuable.

2 Background

The Medical Information Mart for Intensive Care III (MIMIC-III) Dataset is a valuable asset for researchers in the healthcare space. It contains deidentified Intensive Care Unit (ICU) data from patients at Beth Israel Deaconess Medical Center. This dataset has been the source of many state of the art projects, including work being done to predict diagnoses based on clinical notes (Nuthakki et al., 2019), as well as named entity recognition (Kraljevic et al., 2019). Many groups have attempted to use this dataset to model readmission risk. Some have applied Natural Language Processing techniques to the free-text clinical notes as a means to draw insight from this data. Recent advances in transformer based models such as BERT offer a novel approach to use of clinical notes. However, generic BERT was not trained on a clinical ontology and as shown above, would be difficult for a baseline BERT model to comprehend. As such, groups have attempted to fine-tune base BERT models in various clinical contexts. Lee et al. (2019) pretrains BERT on PubMed text, which slightly improves results on tasks such as relation extraction and question answering. More relevantly, Alsentzer et al. (2019) uses BioBERT for pretraining and then trains on MIMIC-III in order to get a higher-performing Bio-ClinicalBERT. There is also ClinicalBERT, trained by Huang et al. (2020) in parallel to Alsentzer, but it does not use BioBERT.

Huang's group focused on a classification task of predicting hospital readmission within a 30-day window. Their main chosen metric is area under the ROC curve, or AUROC, which they obtain 0.672 for on average for notes up to 72 hours, and 0.714 for discharge admissions. For the reasons mentioned above, hospitals find it more valuable to predict readmission on initial notes rather than status as of discharge, so the focus of this paper will be on clinical notes up to 72 hours. Because the clinical notes are so large, and BERT allows for 512 tokens per "sentence", the notes must be chunked. After model prediction, Huang implements a naïve probability prediction that uses an arbitrary constant to adjust for note length, with the supposition that longer notes would mean a higher probability of readmission.

Another state-of-the-art model was built by Rajkomar et al. (2018), which achieved a 0.76 AUC on full clinical notes, but not from the MIMIC-III dataset, so their results cannot be used as a baseline. Their algorithm also is not publicly available nor do they mention use of BERT, but the assumption would be that clinical notes are not too dissimilar among hospitals.

2.1 Significance

This work presents an improvement above current state of the art models in the task of predicting 30-day readmission. In addition to overall performance, the data preparation and model training was performed on clinical notes that were partitioned in hour increments after a patient's initial admission (T+8h, 12, 18, 24, 48, 72). This change in data preprocessing results in an improvement above other models that were built on data partitioned based on the number of days post admission. The more fine-grained breakdown would allow clinicians to understand how readmission risk changes over the course of a patient's stay in the hospital. Additionally, this approach would allow for predictions to be performed right before a "shift change" which is when many care teams meet to discuss the care plan for a patient during the next shift.

3 Methods

3.1 Data

The MIMIC-III dataset consists of 58,976 anonymized hospital Intensive Care Unit (ICU) admissions data from 46,520 unique patients at Beth Israel Deaconess Medical Center. Each admission has a number of free-text clinical notes associated with it. In total, there are 2,083,180 unique clinical notes, a wealth of free-text clinical data. This data, along with associated clinical information such as lab results, diagnoses and procedures are made available to researchers in a set of relational tables. All researchers accessing this data went through HIPAA and privacy training and submitted an application for access.

The goal of this project was to predict unplanned 30-day readmissions. To begin preparing the labels, it is necessary to identify whether or not each individual patient was ever readmitted, then identify if that admission was planned or not, and the time delta between the discharge date and the readmission date was less than 30 days. Any admissions which were for births and any admissions which resulted in death were removed, as these were not the focus of the research.

Next, the clinical notes are processed. For each admission, subset of clinical notes are obtained that were taken at any point between the patient's initial admission and the hour-based time cutoff in question. After merging the time-bounded notes data with the admissions data, all the clinical notes for each admission are concatenated, then some idiosyncratic punctuation and formatting are cleaned. BERT models can only handle a maximum of 512 tokens per sequence, these clinical note documents are quite long, so a decision had to be made on how to break up each document. Huang identified that 318 input words is an ideal input length for ClinicalBERT, given that the tokenizer will on average create about 1.6 tokens per word, and it is optimal to maximize the length of sequences for each prediction. As such, one dataset for each post-admission time interval of interest was created. In addition, the base model uses notes that had all digits, symbols, and punctuation removed. Pre-BERT tokenization for the base model was done using 200, 250, and 350-token divisions, the last of which resulted in 459 BERT tokens on average per note slice.

Finally, to ensure that when the train-test split is created, the split occurs on the admission ID, keeping all note-chunks relating to any given admission in the same dataset. Additionally, because only about 6% of admissions ever result in a readmission, the decision was made to balance the training dataset by downsampling the non-readmitted examples, while retaining an unbalanced, realistic test set. The model was trained on the dataset which contained all clinical notes up to 72 hours after admission. Because notes do not contain the same number of BERT chunks, the label distribution for each chunk does not fit a 50:50 split exactly. Table 1 shows the resultant splits.

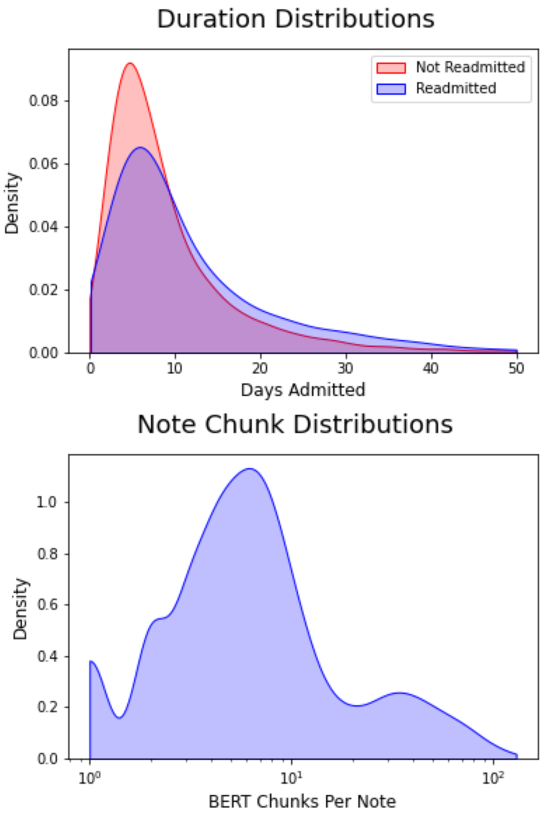

Figure 2: (Top) Distribution of admission duration by readmittance. (Bottom) Training distribution of the number of sequence-chunks by admission, with the median around six.

Dataset	Positive Label	# Records
Train	55.06%	52,363
Validation	52.30%	6,981
Unbalanced Test	7.91%	95,542

Table 1: Population of labels by split.

3.2 Evaluation Metrics

Although F1 score is typically used as a scoring metric for models, many of the papers in this do-main use AUROC, which compares the true positive rate to the false positive rate. To directly compare results with other papers, AUROC will be the main metric. Huang also uses RP80, a metric that measures the recall at a precision of 80%. RP80 is a valuable metric for ensuring high precision and reducing the risk of false alarms, especially important in a clinical setting where care givers often suffer from ”alarm fatigue”.

The individual predictions of the BERT chunks get grouped by the original admission ID, and then have the following formula applied:

$$P(readm = 1|h_{patient}) = \frac{P^n_{max} + P^n_{mean} n/c}{1 + n/c}$$

where the n is number of chunks and c is an arbitrary constant that helps to minimize the impact of longer notes. While Huang chose c = 2 arbitrarily, other values were tested but no change had any significant impact, so the value was kept at 2.

3.3 Base Model

With the prepared data, it is loaded to hardware provided by Google Colab Pro, which provides access to 25 GB of RAM and 16 GB of GPU RAM, important for determining batch size in Pytorch. Using BERT base-uncased on the base model data described in the previous sections, 5 epochs were run. The default learning rate is 5x10^{-5}, but when 4x10^{-4} and 5x10^{-5} were attempted, the training loss would not decrease, so learning rate was reduced to 2x10^{-5}.

Table 2 displays the test measures per level of tokenization, showing that a finer tokenization level results in an increased area under the ROC curve (AUROC) as well as a measure of recall at precision of 80% (RP80).

BERT	Tokens	Loss	AUC	RP80
base-uncased	200	0.27	0.634	0.0479
base-uncased	250	0.29	0.622	0.0308
base-uncased	350	0.34	0.614	0.0100
Bio-Clinical	200	0.21	0.646	0.0137

Table 2: Base model results. Loss shown is for training.

As addendum to the base BERT model, using Bio-ClinicalBERT with 200 tokens in 5 epochs improved the AUROC markedly, but decreased the RP80.

3.4 BERT Optimizations

As stated previously, Huang recommends using 318 tokens because the average BERT token length approaches the maximum 512. Bio-ClinicalBERT was pretrained without removing numbers and normal punctuation, so this tokenization method is used. Because Alsentzer’s Bio-ClinicalBERT is available publicly via Hugging Face, it will be used it as the pretrained model. This results in an average BERT token length of 550 per chunk, with 1220 maximum. Truncation on the first 512 tokens occurs for lengths over 512, and zero-padding for lengths under 512. Two different truncation methods were used, using the first 512 tokens per chunk as well as using the first 256 and last 256 tokens, which was empirically found to be more accurate for some classification tasks (Sun et al., 2020). A smaller learning rate of 1x10^{-5} was tested on the base model and found to have similar results, but is safer to use in general in case the minimum is overshot, so the learning rate was adjusted.

An adjustment to epochs was also tested. Using ten epochs overfits the data such that the training AUROC is maximized close to one, but this model performs poorly on test data, so it should be used.

3.5 Clinical Note Aggregation

To attempt further improvement, the probability aggregation function for admission notes can be adjusted using neural networks. The logit outputs are used because the hidden state training data required over 100 GB of storage. There is no way to hold hidden states in memory without batching; batching cannot continue because the chunks are shuffled and cannot be combined by admission ID.

The logits, labels, and tokens are each appended to a separate array. To get the correct indices, a lookup is performed on the tokens in their known order, using the **index()** function with the newly created token array. With the correct indices, the logits and labels are matched to the admission IDs, effectively undoing the batch shuffling. Logits get grouped by the admission ID into an array, and the difference is taken to get a summarized value per chunk. The difference of the probabilities are also obtained by using the sigmoid function on the original logits. Because each admission ID has a different number of chunks, the length is standardized at ten chunks, with zero-padding for those with less, and using the first five and last five logit differences (while making sure the chunks are

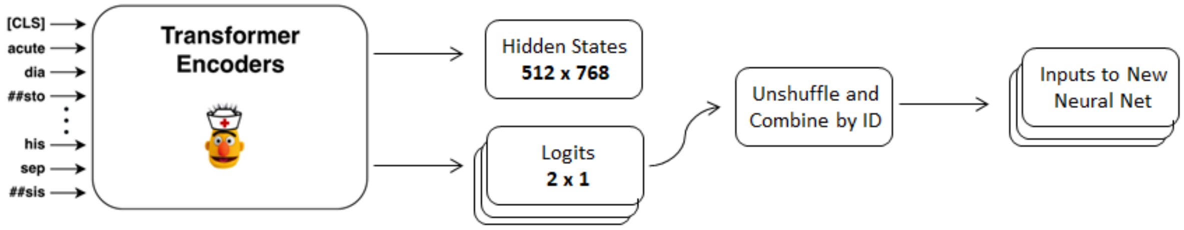

Figure 3: Model with post-BERT processing of data.

in the correct order) for greater than ten chunks. The mean number of chunks was eleven, with a median of six, so ten was chosen as close to the average. Other values for padding were considered but actually performed worse than using zero.

With the arrays of summarized logits and probabilities, the base model is created using a logistic regression model by performing a grid search on parameters, with the C parameter, or inverse of regularization, as the most important. The probability arrays performed better than the logits so, they are carried into different configurations of neural networks. The first configuration contained only dense layers in sequence; the second, a 1-D Convolution Network with kernel size, pooling, stride, and dropout as hyperparameters; the third contained an LSTM with the assumption that previous probabilities should be kept in consideration since some chunks don't actually tell much about the readmission probability. For each, the model goes through a grid search to find the optimal epochs and batch size, and then each optimized model is run 10 times, taking the average score.

4 Results

The results of ClinicalBERT and the post-BERT networks are below. The results are averaged on three different seeds. The best seed had a testing AUC of 0.693 for Bio-ClinicalBERT, while none of the other models approached this value in any seed. The best RP80 was 0.1655, which is in line with Huang's results.

Model	AUROC
Bio-ClinicalBERT	0.685 ±0.006
LogReg	0.674 ±0.002
Dense NN	0.670 ±0.002
CNN	0.669 ±0.003
LSTM	0.671 ±0.003

Table 3: Final results by model type.

Creating a confusion matrix of the best seed shows that the model is quite good at correctly predicting when a patient should be readmitted, but is not so good for non-readmissions. This is preferred because there are relatively few readmissions so getting them correct would be more valuable to the hospital than getting a non-readmission correct.

However, the actual data was undersampled to get the model, so test data with the correct ratio of non-readmits to readmits must be used. The base accuracy for a model should be approximately 0.935, when all predictions are non-readmits. A larger test set is required in order to have more than a few test labels, and admission IDs cannot be duplicated, so a new test set consisting of 1000 non-readmit IDs and 65 readmit IDs is created (the proper ratio may be closer to 950 non-readmits, but 1000 is a good approximation). This data is fed through the best Bio-ClinicalBERT model and the resulting logits and probabilities are also input into the other neural nets.

The results show similar, albeit lower scores to the balanced test dataset, meaning that the models created cannot accurately predict true test data. The best AUROC came once again from the Bio-ClinicalBERT outputs, with an AUROC of 0.665; the logistic regression and neural nets had AUROC of approximately 0.630.

The final method tried was to adjust the cutoff at which something is predicted as a readmission; the idea would be that true readmissions would have a larger spread between their logits per admission ID. So, taking the difference in positive and negative probabilities and finding the mean per admission ID, the threshold probability can be adjusted to above 0.5. Doing so with the training data does not improve anything, while the testing data accuracy slightly improves to 0.940 with the threshold set at 0.9. However, this equates to one extra correct label, not a significant change.

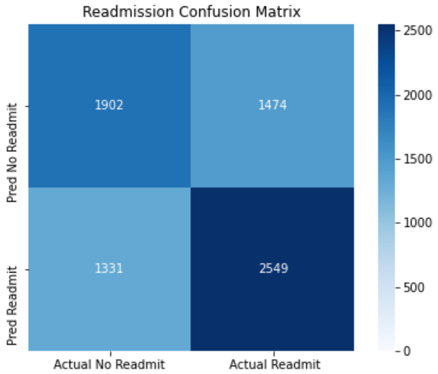

Figure 4: Confusion matrix of the best test results from Bio-ClinicalBERT.

5 Conclusion

This work was able to clean the data and combine Bio-ClinicalBERT with Huang's evaluation method to improve the AUROC.

By tuning Bio-ClinicalBERT on four epochs, 318-token splits, the 256:256 BERT split previously described, a batch size of 8, and a learning rate of 1x10^{-5}, the AUROC score is maximized, just above three standard deviations of the score given in Huang's paper, giving a small improvement. Although the RP80 seems to decrease much further with the neural nets, this misses small changes to the recall expectation that result in large precision gains; for example, the logistic regression on the probability differences results in a test **RP76** of 0.266. As such, this work not only presents an accurate model, but also one that focuses specifically on recall while maintaining high precision, thus minimizing "alarm fatigue" for overburdened clinicians. In short, this work improved upon work done by Huang and others in using a BERT-based model to predict 30-day readmission from free-text clinical notes achieving a 1.5% improvement in AUROC score, as well as a modest improvement in RP80.

The attempts to recombine the ClinicalBERT outputs by admission ID and use them as inputs to another model were not successful in increasing the model score. The scores were not significantly (>5%) lower, but there is no reason to not use the evaluation metric presented by Huang. With larger hardware and more compute time, it may be possible to use the hidden layers as inputs. It also may be worth using training data that has the correct ratio.

Future work should involve improving upon the process for classifying long documents. attempted to improve upon the equation presented by Huang by adding an additional neural network on top of the final hidden layer output from the model, but were unable to find performance gains. Finally, interpretability and explainability of a model are essential to gain trust in a high-consequence clinical setting. As such, future work should leverage attention mechanisms and other techniques to provider care teams with an actionable explanation behind each prediction. This would further solidify this tool as a valuable asset for hospitals.

References

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

A. Hosseinzadeh, Masoumeh T. Izadi, A. Verma, Doina Precup, and D. Buckeridge. 2013. Assessing the predictability of hospital readmission using machine learning. In *IAAI*.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2020. Clinicalbert: Modeling clinical notes and predicting hospital readmission.

Zeljko Kraljevic, Daniel Bean, Aurelie Mascio, Lukasz Roguski, Amos Folarin, Angus Roberts, Rebecca Bendayan, and Richard Dobson. 2019. Medcat – medical concept annotation tool.

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

Siddhartha Nuthakki, Sunil Neela, Judy W. Gichoya, and Saptarshi Purkayastha. 2019. Natural language processing of mimic-iii clinical notes for identifying diagnosis and procedures with neural networks.

Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M. Dai, Nissan Hajaj, Michaela Hardt, Peter J. Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, and et al. 2018. Scalable and accurate deep learning with electronic health records. *npj Digital Medicine*, 1(1).

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2020. How to fine-tune bert for text classification?

S. Sushmita, Garima Khulbe, A. Hasan, Stacey Newman, P. Ravindra, S. Roy, M. D. Cock, and A. Teredesai. 2016. Predicting 30-day risk and cost of "all-cause" hospital readmissions. In *AAAI Workshop: Expanding the Boundaries of Health Informatics Using AI*.

H. Wang, Z. Cui, Y. Chen, M. Avidan, A. B. Abdallah, and A. Kronzer. 2018. Predicting hospital readmission via cost-sensitive deep learning. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 15(6):1968–1978.

Cao Xiao, Tengfei Ma, Adji B. Dieng, David M. Blei, and Fei Wang. Readmission prediction via deep contextual embedding of clinical concepts.