# Transformer for Prediction of Patient Trajectories from Electronic Health Records

**David Huang**
Department of Computer Science
Stanford University
huangdh@stanford.edu

**Prof. David Kim, M.D., Ph.D.**
Department of Emergency Medicine
Stanford University
davidak@stanford.edu

## Abstract

Forecasting future clinical volume could lead to better patient outcomes by improving resource allocation and hospital operations. In this report, we develop a deep learning model for the prediction of patient trajectories using past electronic health record data. We develop a novel embedding strategy that maps patient trajectories to a shared latent space, containing information about diagnostic codes, procedural history, demographics, and visit timing. We explore pretraining a BERT-Base model using masked language modeling. We then finetune our model to predict when a patient will revisit the hospital and what diagnosis category they will present with. Our best performing models achieved top-1 accuracy/F1 scores of 0.35/0.33 for next visit diagnosis category prediction 0.48/0.26 for next visit timeframe prediction. We analyze BERT self-attention weights to interpret how our model makes its predictions, which could lead to clinically-useful insights.

## 1 Introduction

Forecasting aggregate clinical volume is essential to proper resource allocation and staffing [11]. For example, [3] demonstrated that during the COVID-19 pandemic, patients treated during periods of high demand had a two-fold increase in COVID-19 mortality compared to patients treated in periods of low demand. In general, overburdened hospitals have worse patient outcomes for many different conditions [13]. In addition to the strain put on the healthcare system by COVID-19, emergency department (ED) crowding has also been associated with decreased clinical performance [17] and risks to patient safety. In fact, over 90% of hospital EDs report operating at or over capacity [17]. The solution to this problem would involve the ability to effectively predict patient demand patterns at multiple temporal and spatial scales. Predicting future patient trajectories could improve resource allocation and hospital operations to enable better patient outcomes.

Our objective is to develop a model for the prediction of individual patient trajectories, including the timing of their next visit (e.g. next week, next month) and the diagnosis type presented (e.g. circulatory, infectious, injury, etc.). This model could be used alongside previously developed models for predicting patient volume. While previous models have focused on using aggregate statistics [15][10][16], our model could provide more detailed information about when specific patients will revisit and the associated category of diagnosis.

The input to our model is a sequence of a single patient's past ICD-10 diagnostic codes, procedural codes, demographic information, and inter-visit lengths, all of which we refer to as a patient trajectory. We pretrain a Bidirectional Encoder Representations from Transformers (BERT) model using masked language modeling (MLM) [7]. BERT has generated state-of-the-art performance on many natural language processing (NLP) benchmarks [7]. We then finetune our BERT model for multi-task prediction of next visit diagnosis category and timeframe.

CS271: Artificial Intelligence in Healthcare, Autumn 2021, Stanford University

By using a greater variety of input features (e.g. patient demographics, inter-visit lengths) and employing multi-task learning (i.e. predicting both next visit diagnosis and timeframe), we developed a new approach for predicting future patient trajectories. We found that pretraining significantly improves performance for next visit disease category predication. And we found that multitask learning slightly improves next-visit timeframe prediction. We also analyze BERT self-attention weights in order to interpret our model's predictions. Our results demonstrate the feasibility of our approach to forecast detailed information about individual patient trajectories using a BERT model trained on medical records.

## 2 Related work

### 2.1 Univariate and Multivariate Analysis for Patient Forecasting

A variety of approaches have been used for predicting aggregate patient demand. Early work by [15] addressed the problem of forecasting hospital bed requirements. They were one of the first applications of autoregressive integrated moving average (ARIMA) models to this problem, which demonstrated improvements over simple curve fitting for these time-series analyses. [15] has the limitation of being a univariate model (i.e. only using data on the number of hospital beds in use). [10] demonstrated that a multivariate time-series model, integrating information such as laboratory and radiography orders, provided more accurate estimates of ED patient volume compared to a univariate model. [16] presented a non-parametric approach using random survival forests [9] to predict ICU bed occupation. While these studies have used aggregate clinical statistics to forecast patient volume, recent advances in deep learning for individual patient-level prediction tasks might provide a novel method for estimating aggregate patient demand, which we demonstrate in this report.

### 2.2 Deep Learning for Readmission Prediction

Deep-learning prediction models based on individual patient trajectories have recently shown promising results. Deep learning approaches have outperformed many traditional clinically-used prediction models on a range of tasks, including in-hospital mortality, 30-day unplanned readmission, and prolonged length of stay [1]. For instance, [2] evaluated recurrent neural network (RNN) architectures with soft-attention to predict 30-day readmission based on EHR data. They reported an AUROC of 0.739 and an average precision of 0.331. Similarly, [5] used an XGBoost [4] model to predict 72h readmission using EHR data. They reported an AUROC of 0.747 and average precision of 0.233. While current models focus solely on readmission in a fixed timeframe, our model will output predictions across multiple discretized timeframes along with predictions about the disease category. This type of multi-task learning has shown to improve data efficiency, reduce overfitting through shared representations, and improved performance by leveraging auxiliary information [6]. While predicting more detailed information is more challenging, it could provide significantly more useful information for hospital resource allocation and staffing.

### 2.3 Transformers for Disease Prediction

Recent efforts to apply the Transformer architecture to EHR data have focused solely on disease prediction using ICD codes. [18] present a model (BEHRT) that is trained on 1.6 million structured patient records. They pretrain the model using MLM and finetune for the prediction of 301 conditions in future visits. They produced an 8-13% improvement in precision scores compared to previous state-of-the-art models such as Deepr [14] and RETAIN [8]. [12] also developed a Transformer model (Med-BERT) for disease prediction using structured EHR data. They employ an additional pretraining task of prolonged length-of-stay prediction as well as a larger number of patient records (20 million). Both models achieve state-of-the-art performance for disease prediction from structured EHR data. We plan on improving upon these models by incorporating significantly more features into our embeddings, including procedural codes, patient demographics, and inter-visit lengths. In addition, we plan on simultaneously optimizing for the task of next visit timeframe prediction, which could integrate information about lengths between visits into the learned embeddings, which is a notable difference between structured EHR data and free-text data.

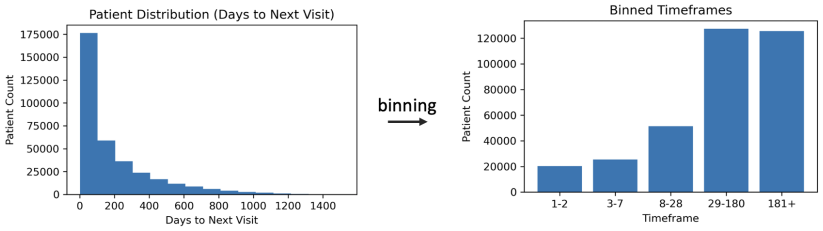


Figure 1: Binning strategy for next visit timeframes. Each patient trajectory was labeled with the timeframe of the next visit, which was discretized into the five bins shown on the right.

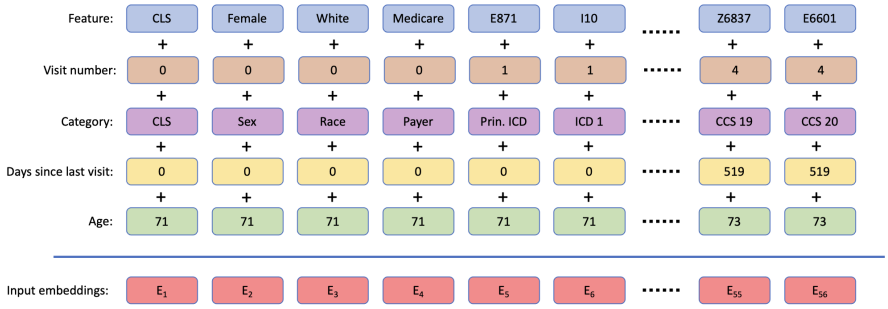


Figure 2: Embedding strategy. A patient trajectory is converted into a sequence of embeddings that capture medical, temporal, and demographic information.

## 3 Data

Our dataset is provided courtesy of Prof. David Kim in the Stanford Department of Emergency Medicine. Our unlabeled pretraining dataset consists of 59 million emergency department (ED) visits, representing the trajectories of 18 million adult patients located in California, Florida, and New York between October 1, 2015 and December 31, 2018. Each de-identified visit record includes patient demographics (age, sex, and race), visit characteristics (primary payer, length of stay, and days since last visit), up to 24 ICD-10 diagnosis codes, and up to 20 procedural codes. The dataset was previously filtered to include patient trajectories that consist of three or more documented visits. In our dataset, the average age ($\pm$ SD) was $45.85$ ($\pm 23.12$), and the average number of visits in a patient trajectory ($\pm$ SD) was $5.81$ ($\pm 16.71$).

For our labeled finetuning dataset, we truncated trajectories before their final visit, such that we could use the diagnosis at the final visit and the number of days until the final visit as labels. The category of diagnosis at the final visit was determined using the single principle ICD-10 code, which was binned into 14 categories: neurologic, infectious, gastrointestinal, injury/poisoning, circulatory, obstetric/gynecologic, respiratory, musculoskeletal, dermatologic, endocrine/FEN/immune, neoplastic, psychiatric/substance, genitourinary, and other. The number of days until a patient's final visit was widely distributed, with most final visits falling in the 1-200 range, as shown in Figure 1. Therefore, we binned final visit timeframes into 5 categories: 1-2 days, 4-7 days, 8-28 days, 29-180 days, and 180+ days. We hypothesized that readmission in 1-2 days, 4-7 days, and 8-28 days would be clinically distinguishable, which is why we split them into separate bins instead of using a single bin for 1-28 days. In our dataset, the average number of days until a patient's final visit ($\pm$ SD) was $80.66$ ($\pm 148.02$).

# 4 Approach

## 4.1 Embedding Strategy

Features within patient trajectories (e.g. age, race, principle diagnosis for the first visit, principle procedural code for the third visit, etc.) were mapped to 288-dimensional embedding representations, similar to mapping words to word embeddings in traditional NLP. Given the fixed input length of BERT, patient records with less than 128 features were padded to length 128 and records with over 128 features were truncated to only the last 128 features, since recent diagnoses are assumed to be more important. Pad tokens were masked from the attention and loss calculations. The average length of an input sequence ($\pm$ SD) was $27.23$ ($\pm 49.32$) tokens.

Concretely, the sequence of input embeddings to our BERT model, $E_1, E_2 ... E_{128}$, is the sum of feature embeddings, visit number embeddings, category embeddings, days-since-last-visit embeddings, and age embeddings (Figure 1). Feature embeddings correspond to the patient's sex, race, primary payer, ICD-10 diagnostic codes for each visit, and CCS procedural codes for each visit. Each input trajectory always starts with a CLS token (for extracting pooled embeddings), a token for the patient's sex, a token for the patient's race, and a token for the patient's primary payer. Then, tokens for a patient's ICD diagnosis codes and CCS procedural codes are grouped by visit and ordered from earliest visit to most recent visit. Within a visit, diagnosis codes are ordered with the principal diagnosis first. Since there are variable numbers of codes for each visit, we need to use position embeddings to denote the visit number. Thus, instead of using standard BERT position embeddings (described in [7]), we use the visit number to indicate the position of each diagnosis or procedural code in the patient trajectory. We then add category embeddings, which correspond to the category of the corresponding feature (i.e. principle or secondary ICD diagnosis). Finally, embeddings corresponding to the number of days since the patient's last visit and the patient's age for that visit are added in to provide temporal context. Incorporation of inter-visit length embeddings has not previously been explored and might improve the interpretation of a patient trajectory.

## 4.2 BERT Pretraining

We pretrained a BERT-Base model to learn EHR feature embeddings. The basis of the BERT model is the multi-headed self-attention layer, which assigns each input token ($x$) unique query ($Q_j$), key ($K_j$), and value ($V_j$) embeddings. Output embeddings are computed by

$$a_j = softmax(\frac{Q_j(x)K_j(x)^T}{\sqrt{d_c}})V_j(x) \tag{1}$$

where $a_j$ is the attention-weighted output of the jth attention head and $d_c$ is the embedding dimension. In multi-headed attention, the outputs are combined, enabling each attention head to attend to different areas of the input sequence. We pretrained BERT using masked language modeling (MLM), where the model learns to predict a masked token using the sequence context. We masked with a probability of 15% (80% mask, 10% randomly permute, 10% unchanged). We added a MLM head that converts the output embedding sequence into a sequence of vocab-size length embeddings, which models a distribution of token probabilities. We use a cross-entropy loss function with non-masked tokens ignored from the loss calculation. Our MLM strategy was inspired by [7]. We used a BERT-Base model retrieved from the HuggingFace library.[1]

## 4.3 BERT Finetuning

We finetuned our model for next visit diagnosis category prediction and next visit timeframe prediction. For our finetuned BERT model, we used the same architecture from pretraining, but we replaced the MLM head described above. Instead of using the sequence output from BERT-Base, we used pooled output embeddings, which correspond to the CLS embedding at the beginning of the patient trajectory (Figure 2). Pooled embeddings were used as input into two prediction heads for either next visit diagnosis category prediction or next visit timeframe prediction. Each prediction head consisted of a dropout and fully-connected layer with a cross-entropy loss function. For the timeframe prediction

[1] https://github.com/huggingface/transformers

head, we weighted individual loss values by the fractional occurrence of the most common timeframe bin divided by the fractional occurrence of the given timeframe bin in order to account for dataset imbalance. For our multitask model, we trained both prediction heads simultaneously by adding together the cross-entropy losses. We used a 90/5/5 train/validation/test split. We experimented with six different finetuning methods to evaluate the contribution of pretraining and multitask learning.

# 5 Experiments/Results/Discussion

## 5.1 Pretraining

We pretrained using MLM for 1 epoch over 4 million examples with a batch size of 28 and an Adam optimizer. Our model consisted of 6 hidden layers, each with an encoder size of 288 and feed-forward size of 512, as used in [18]. We used a Gaussian Error Linear Unit (GELU) activation and dropout rate of 0.1, as described in [7]. We used a learning rate of 1e-4 and a linear schedule with warmup, as suggested by [7]. We obtained a final cross-entropy loss of 2.52 and perplexity of 12.45.[2] Due to compute constraints, we needed to end pretraining early. As a result, the model did not see all 18 million unlabeled examples.

## 5.2 Comparing Performance of Our Models

We evaluated six models for prediction of next visit diagnosis category and/or timeframe (Table 1). For each model, we finetuned for 10 epochs using a batch size of 512. We used the same model architecture from pretraining, but we increased the dropout rate to 0.3 to prevent overfitting and lowered the learning rate to 1e-5. All our models performed significantly better than a random classifier, which would have a top-1 accuracy of 0.07 for diagnosis category prediction and 0.36 for timeframe prediction by simply predicting the most common class.

| | | Diagnosis category | | Timeframe | |
|---|---|---|---|---|---|
| Model Prediction Task | Pretrained | Top-1 Accuracy | F1 Score | Top-1 Accuracy | F1 Score |
| Multitask | No | 0.26 | 0.20 | 0.47 | 0.25 |
| Multitask | Yes | 0.34 | 0.32 | **0.48** | **0.26** |
| Diagnosis category only | No | 0.29 | 0.25 | - | - |
| Diagnosis category only | Yes | **0.35** | **0.33** | - | - |
| Timeframe only | No | - | - | 0.46 | 0.22 |
| Timeframe only | Yes | - | - | 0.47 | 0.24 |

Table 1: We trained and evaluated six model variations on a held-out test set. We evaluated our models using accuracy and F1 scores from our models' top predictions (our model output is a probability score for each category). Metrics for our best-performing models are highlighted.

We found that pretraining led to a significant increase in performance for diagnosis category prediction and a small increase in performance for timeframe prediction (Table 1). Concretely, the top-1 accuracy and F1 scores corresponding to diagnosis category for the pretrained multitask model were 0.34/0.32, which were significantly higher than the scores of 0.26/0.20 for the non-pretrained multitask model. This is likely because MLM closely mimics the task of diagnosis category prediction. In fact, most of the features in a patient trajectory are ICD-10 codes, so MLM primarily involves predicting information about diagnosis types. For timeframe prediction, due to pretraining, there was a small increase in top-1 accuracy and F1 score from 0.47/0.25 to 0.48/0.26 for our multitask model. This is likely because MLM does not involve temporal prediction (i.e. for inter-visit lengths), so predicting the time until a future visit is a substantially different task.

We also found that multitask learning led to a small advantage for next visit timeframe prediction, since the top-1 accuracy and F1 performance of the multitask model was 0.48/0.26, which is slightly higher than the performance of 0.47/0.24 for the model trained only on next visit timeframe prediction. One way to rationalize this is that there is some association between the time to a patient's next visit and the category of a patient's next diagnosis. In Figure 3, we show that there is a correlation between

[2]Perplexity is a standard metric for language models equal to the exponential of the cross entropy loss.

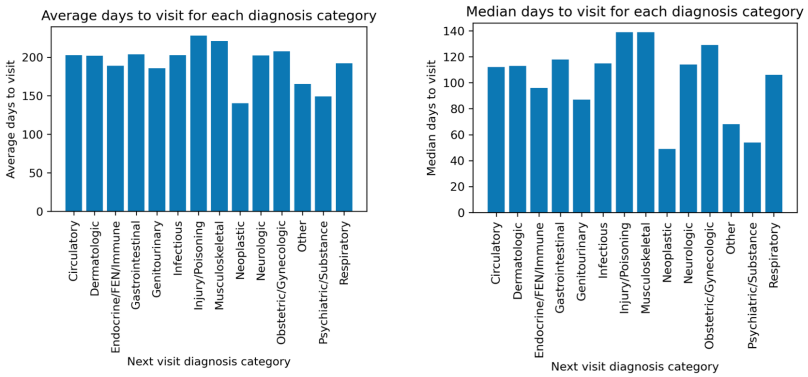


Figure 3: Association between the time until a future visit and the category of diagnosis at that visit.

the diagnosis a patient receives at a future visit and the time until that visit. In our dataset, patients who will have a neoplastic diagnosis will revisit in an average/median of 140/49 days. But patients who will have an injury/poisoning diagnosis will revisit much later in an average/median of 228/139 days. This makes sense because severe neoplastic diagnoses (i.e. malignant tumors) are very likely to be followed by future hospital visits. On the other hand, injury/poisoning diagnoses might occur more often with younger, healthier patients who don't revisit as soon. Thus, there might be some shared information in the embeddings learned from predicting disease category that have led to the small increase in performance on timeframe prediction for our multitask model.

On the other hand, multitask learning did not benefit disease category prediction, likely because it is a much easier task compared to timeframe prediction. For instance, our binned timeframes, although motivated by clinical relevance, are somewhat arbitrary. A patient who revisits in 1-2 days, for example, might have no distinguishing features from a patient who revisits in 3-7 days. Therefore, pairing timeframe prediction with disease prediction might introduce unnecessary noise that slightly reduces our multitask model's performance.

### 5.3 Examining Our Model Performance

In Figure 4, we show performance metrics on a held-out test set for our pretrained multitask BERT model. We calculate category-specific ROC curves. If the output of our model for a single example is $y'$, we use only the $y'_i$ values to calculate our ROC curves, which corresponds to the model's prediction for the $i$th category. Below, we also show confusion matrix heatmaps by comparing our model's predicted category with the ground truth category.

For next visit category prediction, our model performed best on predicting future obstetric/gynecologic (AUROC = 0.98), neoplastic (AUROC = 0.89), and psychiatric/substance (AUROC = 0.87) diagnoses (Figure 4). One reason for this is that these types of diagnoses tend to follow a predictable pattern. For instance, diagnoses within obstetrics/gynecology commonly relate to pregnancy, which can be associated with multiple hospital visits, making future diagnoses more predictable. Similarly, psychiatric/substance diagnoses could commonly be associated with multiple revisits for the same underlying condition, which makes prediction of future visits easier. Our model performed worst on prediction of future injury/poisoning (AUROC = 0.65), infections (AUROC = 0.66), and gastrointestinal (AUROC = 0.70) diagnoses. This makes intuitive sense because these types of diagnoses usually correspond with single hospital visits, which makes predicting these diagnoses very difficult.

For next visit timeframe prediction, we observed relatively lower performance overall (Figure 4). As discussed earlier, the factors behind this might include the arbitrary binning strategy and poor applicability of MLM as a pretraining task for timeframe prediction. Our model performed best on predicting future visits in the 180+ range (AUROC = 0.73). This could be from the model's understanding of a reasonably healthy versus unhealthy patient, as a healthy patient would likely revisit in 180+ days. Our model performed worst on predicting the 29-180 day range (AUROC = 0.57). This was also the most common label in our unbalanced dataset (Figure 1). Thus, our model

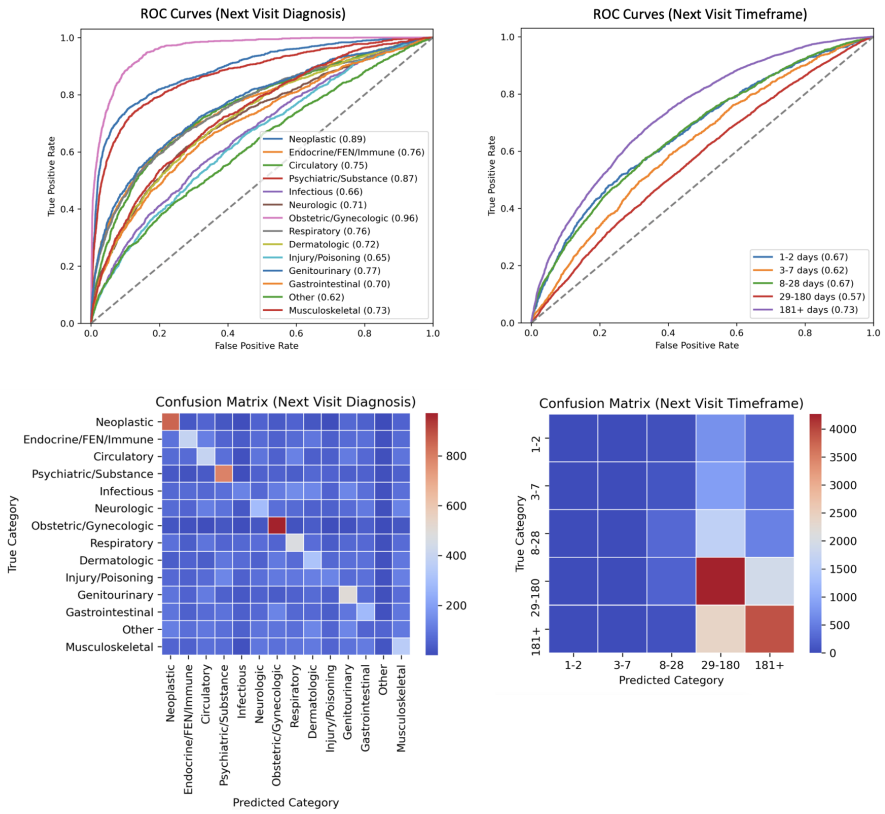


Figure 4: Receiver operating characteristic (ROC) curves and confusion matrix heatmaps for our pretrained multitask BERT model. Next visit diagnosis category prediction is shown on the left and next visit timeframe prediction is shown on the right.

might be simply predicting the most common label, leading to the abundance of false-positivies for the 29-180 day timeframe.

### 5.4 Analysis of BERT attention

In Figure 5, we show BERT attention weights for all 12 attention heads for the last layer of our model on an example that was correctly classified (Figure 5, left) and an example what was wrongly classified (Figure 5, right). Although we labeled the vertical axis with the input features, the actual embeddings are a sum of multiple embedding types, as described in Figure 1. Generally, we see that our model attends more heavily to diagnostic and procedural codes in later visits (greater intensity near the bottom of the heatmaps), which is reasonable since more recent visits are likely better indicators of future visits.

For the correctly predicted example (Figure 5, left), our model identified future neoplastic diagnosis in 8-28 days. To make this prediction, the attention heads focused on ICD-10 codes for secondary malignant neoplasms, cerebral edema, and visual hallucinations. In particular, the embeddings for malignant neoplasms were highly attended to, which clinically makes sense, as patients with malignant neoplasms (i.e. tumors) are very likely to revisit for the same condition. In addition, embeddings for cerebral edema and visual hallucinations were strongly attended to, although they are not in the neoplastic category. These features might indicate that this patient is in critical condition, which could inform the model's prediction of a closer timeframe (8-28 days) to their next visit.

For the incorrectly classified example (Figure 5, right), our model predicted that the next visit would be a gastrointestinal diagnosis in 180+ days, when the true label was an infectious diagnosis in 8-28 days. Compared to the correctly classified example, these attention weights were more scattered

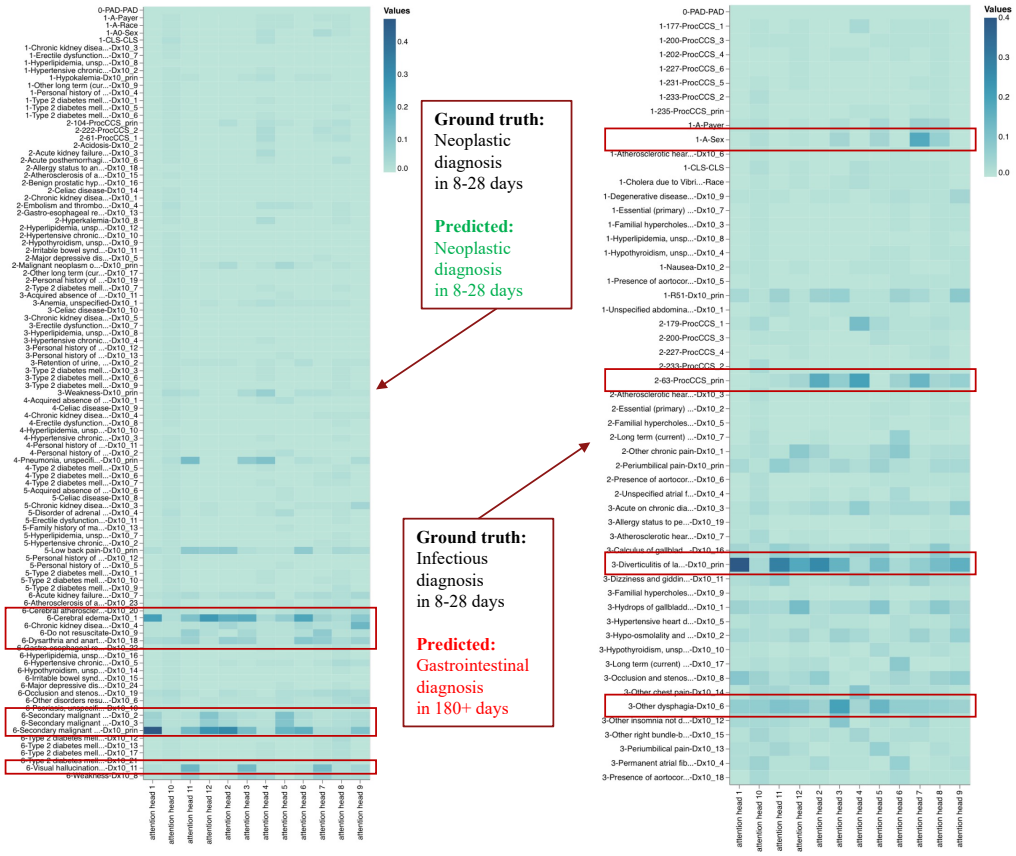


Figure 5: BERT self-attention weights for a correctly predicted example and wrongly predicted example. Red boxes highlight feature embeddings that were likely important to our model's prediction.

across the patient trajectory, indicating that there were no clear features in the input to predict future infectious diagnosis in 8-28 days. As expected, our model attended highly to ICD-10 codes for diverticulitis and dysphagia, which are both in the gastrointestinal category. Thus, our model incorrectly uses past diagnoses of gastrointestinal disease to predict a future gastrointestinal diagnosis. And since gastrointestinal diagnoses tend to have longer revisit times (Figure 3), our model could also be using those features to predict the 180+ days timeframe to their next visit. The model also attends to a cardiovascular procedure and demographic information, and there could be information within those embeddings (from Figure 2) that the model is incorporating. This is a particularly difficult example, even for a trained clinician, since there is very little past medical information to indicate future infectious diagnosis, which is why our model performed poorly on this category (Figure 4).

## 6 Conclusion

In this report, we describe an effective strategy that applies principles from NLP to structured EHR data in order to forecast future patient trajectories. We include significantly more features (i.e. procedural, demographic, and temporal information) and data than previously developed models. Our best performing models achieved top-1 accuracy/F1 scores of 0.35/0.33 for next visit diagnosis category prediction and 0.48/0.26 for next visit timeframe prediction. We show that pretraining provides a significant advantage, especially for next visit diagnosis prediction. We also show that analyzing BERT self-attention weights allows for model interpretation, which can lead to clinically-useful insights. Future work should significantly extend pretraining to all 18 million examples for multiple epochs. Future work should also address timeframe class imbalance by using different binning methods and data augmentations. Furthermore, future studies could experiment with different pretraining methods to better match the task of timeframe prediction.

## 7 Contributions

Prof. David Kim provided the datasets and gave high-level guidance on clinical aspects. David Huang developed the code, ran the experiments, performed the analysis, and wrote the paper.

## References


[1] Rajkomar A, Oren E, Chen K, and et al. Scalable and accurate deep learning with electronic health records. *npj Digital Med*, 18(1), 2018.

[2] S. Barbieri, J. Kemp, O Perez-Concha, and et al. Benchmarking deep learning architectures for predicting readmission to the icu and describing patients-at-risk. *Sci Rep*, 10(1111), 2020.

[3] Dawn M. Bravata, Anthony J. Perkins, Laura J. Myers, Greg Arling, Ying Zhang, Alan J. Zillich, Lindsey Reese, Andrew Dysangco, Rajiv Agarwal, Jennifer Myers, Charles Austin, Ali Sexson, Samuel J. Leonard, Sharmistha Dev, and Salomeh Keyhani. Association of intensive care unit patient load and demand with mortality rates in us department of veterans affairs hospitals during the covid-19 pandemic. *JAMA Network Open*, 4(1), 01 2021.

[4] Tianqi Chen and Carlos Guestrin. Xgboost. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

[5] F.P. Chmiel, D.K. Burns, M. Azor, and et al. Using explainable machine learning to identify patients at risk of reattendance at discharge from emergency departments. *Sci Rep*, 11(21513), 2021.

[6] Michael Crawshaw. Multi-task learning with deep neural networks: A survey, 2020.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[8] Choi E and et al. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. *In Advances in Neural Information Processing Systems*, page 3504–3512, 2016.

[9] Ishwaran Hemant, Kogalur Udaya, Blackstone Eugene, and Lauer Michael. Random survival forests. *The Annals of Applied Statistics*, 2(3):841—-860, 2008.

[10] Spencer S. Jones, R. Scott Evans, Todd L. Allen, Alun Thomas, Peter J. Haug, Shari J. Welch, and Gregory L. Snow. A multivariate time series approach to modeling and forecasting demand in the emergency department. *Journal of Biomedical Informatics*, 42(1):123–139, 2009.

[11] Aiken L, Clarke S, and Sloane D. Hospital staffing, organization, and quality of care: cross-national findings. *International Journal for Quality in Health Care*, 14(1):5–14, 2002.

[12] Rasmy L, Xiang Y, Xie Z, and et al. Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *npj Digit Med*, 14(86), 2021.

[13] Phillips M, Sarff L, Banerjee J, Coffey C, Holtom P, Meurer S, Wald-Dickler N, and Spellberg B. Effect of mortality from covid-19 on inpatient outcomes. *Journal Med Virology*, 2021.

[14] Miotto R, Li L, Kidd B, and Dudley J. Deep patient: An unsupervised representation to predict the future of patients from the electronic health records. *Sci. reports*, 26094, 2016.

[15] Farmer RD and Emami J. Effect of mortality from covid-19 on inpatient outcomes. *Journal Med Virology*, 2021.

[16] J. Ruyssinck, J. van der Herten, R. Houthooft, F. Ongenae, I. Couckuyt, B. Gadeyne, K. Colpaert, J. Decruyenaere, F. De Turck, and T. Dhaene. Random survival forests for predicting the bed occupancy in the intensive care unit. *Computational and mathematical methods in medicine*, 2016.

[17] Chase V, Cohn A, Peterson T, and Lavieri M. Predicting emergency department volume using forecasting methods to create a surge response for noncrisis events. *Acad Emerg Med*, 19(5):569–576, 2012.

[18] Li Y, Rao S, Solares JRA, and et al. Behrt: Transformer for electronic health records. *Sci Rep*, 10(7155), 2020.