ClinicNet: Clinical Decision Support and Patient Progression Prediction through Deep Neural Networks

Delaney Sullivan
Computer Science
Stanford University
dsull@stanford.edu

Jonathan Wang
Biomedical Informatics
Stanford University
jonwang1@stanford.edu

Adam Wells
Statistics
Stanford University
adamw356@stanford.edu

Alex Wells
Biomedical Informatics
Stanford University
awells2@stanford.edu

March 23, 2018

Abstract

Medical decision making is delegated almost entirely to the experience of physicians, leading to inconsistency and variability in practice. Our group has developed systems to provide both automated clinical decision support and patient progression predictions through feed-forward and recurrent neural network models. Given a patient's current status based on information in the EHR system, we accurately predict what happens next to the patient in the next 24 hours. We showcase two models, one to replace existing human derived order sets, and another to provide general patient timeline prediction. Our feed-forward model (AUROC: 0.949) outperforms human authored order sets (AUROC: 0.723) in predicting recommendable clinical orders for physicians in 24 hours. Our recurrent neural network model (AUROC: 0.907) performed better than a baseline model based on recommending the top 25 events (AUROC: 0.753) in making patient progression predictions.

1 Introduction and Related Work

Despite the rapid and widely successful incorporation of data-driven decisions into a plethora of different industries, when it comes to healthcare, much of high quality practice and decision making is still dependent almost entirely upon the experience and knowledge of a single physician. Consistency and variability plague much of the current practice. We see this manifested in a variety of scenarios; for example, in one study, 25% of heart attack patients had not been receiving the proper aspirin medication. **Many clinical decisions in the hospital lack evidence-based support**, due in part to the difficulty of performing randomized controlled patient experiments, as well as high variability in compliance to evidence-based guidelines, ranging widely from 20-80% [1]. Moreover, only about 11% of recommendation guidelines are backed by high quality evidence. Thus, the progressively growing amount of literature and instructions required to make appropriate medical decisions is left to the anecdotal and individual experience of physicians [2]. There is simply a lack of knowledge and ability to keep up with the growing amount of skill, information, and communication required to help and heal people most effectively [3]. Recent policies in the United States such as the HiTech act (2009) and Medicare Access and CHIP Reauthorization Act of 2015 have resulted in the widespread adoption of electronic health records (EHRs) in over 80% of hospitals [4, 5]. **The opportunity to feed this information into a learning health system is a promising and potentially life-saving way to form data-driven decisions to support current clinical practice.**

Currently, one form of **clinical decision support** already implemented within hospital systems involves **human authored order sets**. When ordering clinical items for a patient in the hospital, clinicians can search for and choose preformed order sets for common conditions. These order sets are manually developed by clinical committees, a time consuming and unscalable process. Existing literature has shown the ability of probabilistic topic models modeled off of the "Netflix" and "Amazon" recommender algorithms, to outperform preformed order sets in terms of suggestions for doctors in precision, recall, and f1 scores [6]. These methods are not only more accurate than current guidelines, they are also vastly more scalable than manually creating thousands of custom order sets. Recent developments in computing power and the large amount of clinical data available is the perfect setting for artificial intelligence algorithms, like deep neural networks, to provide meaningful and accurate predictions. There has already been great success in applying such tools to tasks such as medical image segmentation and predicting diagnoses, readmission, length of stay, and death [7–9]. Previous studies have used deep-learning models for decision making in healthcare and have shown that deep neural networks perform better than shallower machine learning architectures [10]. Here, we train, develop, and test a set of neural networks to predict clinically meaningful order sets for automated clinical order decision support and patient progression over time.

Our neural network models use patient-level features from the **electronic health record (EHR)** as inputs (Figure 1). We showcase two such models, **ClinicNet** and **ClinicLSTM**. ClinicNet is a feed-forward neural network trained predict which of the 1,452 clinical items that are in existing order sets (e.g a medication) a patient will receive in the next 24 hours. ClinicNet significantly outperforms currently used human authored order sets on this task (**precision: 0.411** vs 0.208, **recall: 0.612** vs 0.476, **AU-**

ROC: 0.949 vs 0.723). ClinicLSTM is a recurrent neural network using LSTM units to predict which of 2,994 patient progression events (e.g admitted to ICU, seen by a cardiologist, etc.) a given patient will experience in the next 24 hours. We find that the ClinicLSTM can predict these events quite well with a **precision of 0.366**, **recall of 0.318** and **AUROC of 0.907**.

Inputs: Feature matrix capturing the patient context within the hospital. We had over 7,500 features including demographic, lab, treatment, medical history, diagnosis, etc. every time an item is ordered for a patient, which attempts to capture the patient status (all information available in the EHR). Outputs: A binary response vector of 1,452 clinical items ordered in the next 24 hours (ClinicNet) or a binary response vector of 2,994 different possible progression events that could occur in the next 24 hours (ClinicLSTM).

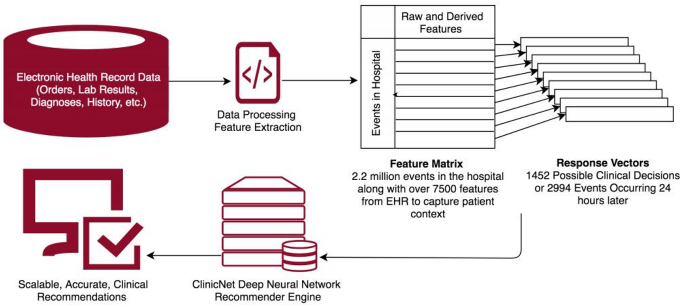
Electronic Health Record Data (Orders, Lab Results, Diagnoses, History, etc.)
Data Processing Feature Extraction
Raw and Derived Features
Events in Hospital
Feature Matrix
2.2 million events in the hospital along with over 7500 features from EHR to capture patient context
Response Vectors
1452 Possible Clinical Decisions or 2994 Events Occurring 24 hours later
Scalable, Accurate, Clinical Recommendations
ClinicNet Deep Neural Network Recommender Engine

Figure 1: Clinic Recommender Process Diagram

2 Dataset and Features

Data was extracted from **Stanford Medicine Research Data Repository (STARR)** with Institutional Review Board (IRB) and hospital approval from Stanford HealthRex Laboratory [11]. In total, the data used for this project spans from 2008-2014 and includes 2.2M patient encounters and 45K distinct clinical items. Clinical items consist of 10,000 medications, 1,600 lab tests, 1,200 images, and 1,000 nurse orders. All medications were grouped according to RxNorm mappings to administration route and active ingredients [12]. Other data elements included demographics, ICD9 diagnoses codes, medical team visits, stays in the ICU, admission, discharge status etc. Pre-processing was done to allow each patient to be modeled as a timeline of clinical item events at discrete time points. Clinical items were removed based on the 80/20 power law distribution, such that any items with less than 256 orders were removed. This allowed for greater computational efficiency at the sacrifice of only 2% of items ordered. Data processing was further performed using a pipeline developed by the Stanford HealthRex Laboratory for building feature matrices from STARR data. We modified and optimized the code to improve its runtime and to narrow down the features in the dataset. Ultimately this resulted in a feature matrix consisting of over 7,500 columns, with corresponding response vectors consisting of 1,452 recommendable clinical items (ClinicNet) or 2,994 patient timeline events (ClinicLSTM). ClinicNet trained on 750,000 rows split into 92/4/4 train/dev/test split with events from the human authored order sets split randomly between the dev and test, and ClinicLSTM was trained on 28,859 distinct 48 hour patient sequences and used a 90/5/5 train/dev/test split.

3 Methods

3.1 Feed-Forward Model - ClinicNet

A deep neural network was trained to automate order recommendations by predicting which clinical items a patient would receive in the next 24 hours. This is a multi-label prediction problem that lends itself well to a feed-forward neural network where the output layer contains sigmoid activation functions that produce a 1,452 x 1 dimensional output, corresponding to probabilities that a patient will receive each of the clinical items. The inputs included over 7,648 features pertaining to patient demographics and past clinical item orders and lab test results. A weighted cross-entropy loss function (1) was used during training to increase the penalty associated with false negative predictions. This was necessary as the large majority of patients only receive a small proportion of the total number of clinical items available, which causes the dataset to have highly unbalanced classes.

$$-1/m\sum_{i}^{m} wy^{(i)}log(\hat{y}^{(i)}) + (1-y^{(i)})log(1-\hat{y}^{(i)}) \qquad (1)$$

Hyperparameters including the learning rate, dropout rate, number of layers, number of neurons per layer, and the weight in the loss function were tuned using a coarse-to-fine approach (Table 1). The development and test sets were created such that they only included instances where an order set was used. This enables us to directly compare the performance of ClinicNet against the existing recommendation system while also ensuring that the distribution of the development and test sets match real-world scenarios in which an order set is used. After testing numerous models on a development set, it was determined that the best performing model had five hidden layers with 500, 475, 450, 425, and 400 neurons, respectively (Figure 2).

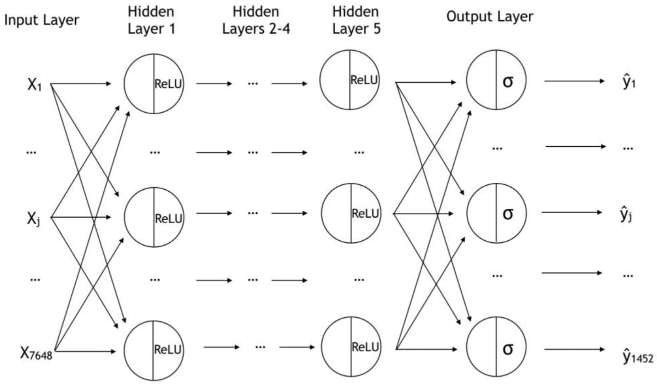

Figure 2: ClinicNet Architecture

3.2 Recurrent Model - ClinicLSTM

A recurrent neural network was trained to predict patient progression in the next 24 hours (ClinicLSTM) using a patient's corresponding medical history extracted from EHR. Each patient in the hospital will often experience many patient progression events in a given 24-hour time period, making this a multi-label prediction problem. To format the input in a manner suitable for an RNN, we grouped the data by patient encounter, resulting in a timeline of events for each patient in the database along with the 9,290 features associated with the patient (e.g prescribed medications, demographics, lab test results, etc.) at each time point. Patient timelines were then split into 28,859 distinct 48-hour segments. We randomly selected the 48-hour time segments of 90% of the patients to place into the training set, 5% into the development set, and 5% into the test set. Hyperparameters including learning rate, number of layers, number of hidden units per LSTM block, and the positive weight in the loss function were tuned using a coarse-to-find approach (Table 1).

ClinicLSTM uses LSTM blocks followed by a dense output layer with a sigmoid activation function after each time-point (Figure 3). The model uses a weighted cross entropy loss function and was trained for 5 epochs (1). After each time-point in a patient's timeline, ClinicLSTM outputs a 2,994 x 1 dimensional vector, corresponding to probabilities that a patient will experience each of the progression events in the next 24 hours.

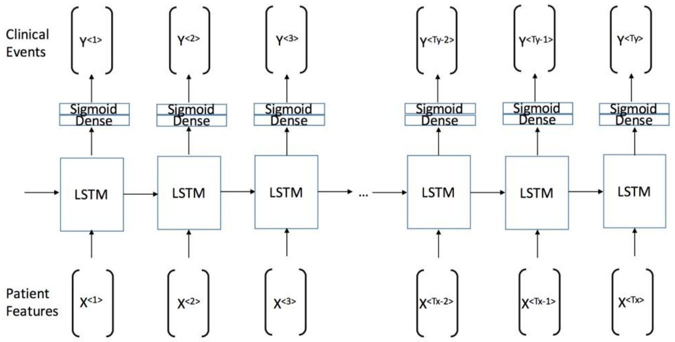

Figure 3: ClinicLSTM Architecture

	ClinicNet	ClinicLSTM
Layers	5	1
Hidden Units per Layer/Cell	500,475 450,425,400	800
Learning Rate	5e-3	1e-3
Dropout Rate	0.05	0
Positive Weight	6	5

Table 1: ClinicNet and ClinicLSTM Hyperparameters

4 Results

4.1 ClinicNet Model Results

The ClinicNet feed-forward model outperformed the order sets based on precision, recall, and AUROC.

Model	Precision	Recall	F1 Score	AUROC
ClinicNet	**0.411**	**0.612**	**0.492**	**0.949**
Order Set	0.208	0.476	0.289	0.723

Table 2: ClinicNet and Order Set Performance

Perhaps most noticeable is ClinicNet's precision score compared to the order sets, indicating that the recommendations the model produced could potentially be treated with much higher confidence by a medical professional.

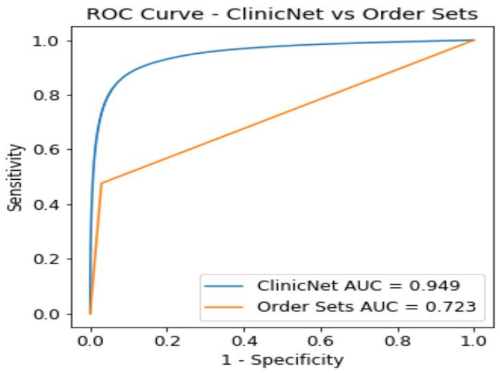

Figure 4: Clinical Item ROC Curve

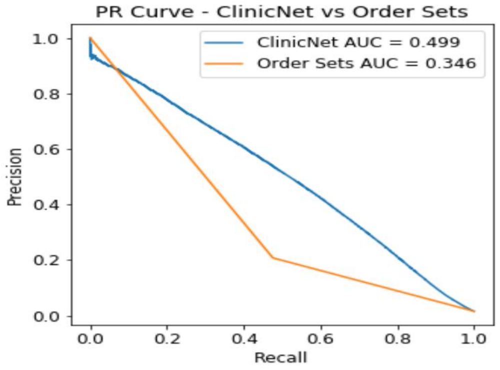

Figure 5: Clinical Item P-R Curve

Furthermore, when comparing on a per clinical item basis, we see that ClinicNet outperformed the Order Sets on 81.7% of the clinical items. The median AUROC improvement was 0.12.

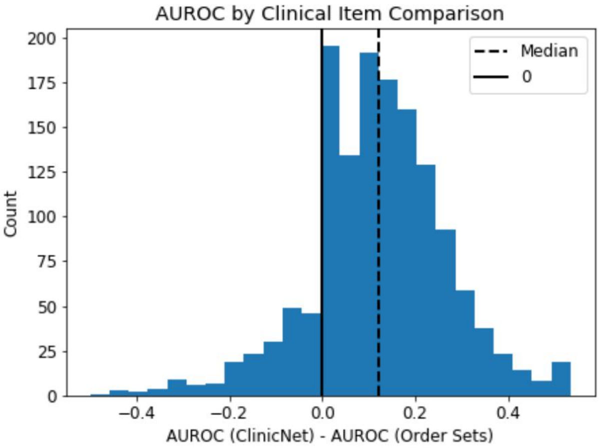

Figure 6: AUROC Difference by Clinical Item

4.2 ClinicLSTM Model Results

We compared the ClinicLSTM prediction results to those of a **"Majority" patient progression model**. For each time-step in a given patient's timeline, the "Majority" model predicts the 25 (median number of events in the next 24 hours over dataset) progression events that have occurred most often in the entire training set. We find that ClinicLSTM outperforms the "Majority" model in both precision, F1-score, and AUROC, while falling short in recall.

Model	Precision	Recall	F1 Score	AUROC
ClinicLSTM	**0.366**	0.318	**0.340**	**0.907**
Majority	0.140	**0.535**	0.221	0.723

Table 3: ClinicLSTM and "Majority" Performance

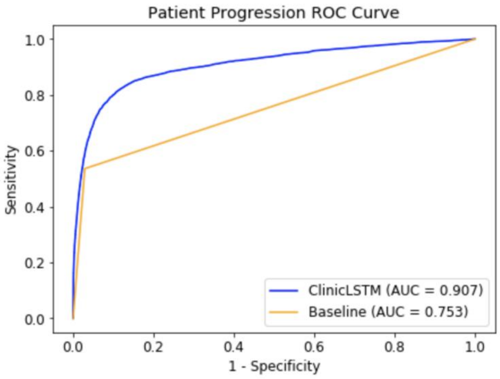

Figure 7: Patient Progression ROC Curve

5 Discussion

Our model provides effective clinical decision support for physicians in a scalable and data-driven manner. Current methods for curating order sets are time consuming, unscalable, and not specific to a given patient. By leveraging data from the EHR, we demonstrate how deep learning can be used to improve upon this current standard of care.

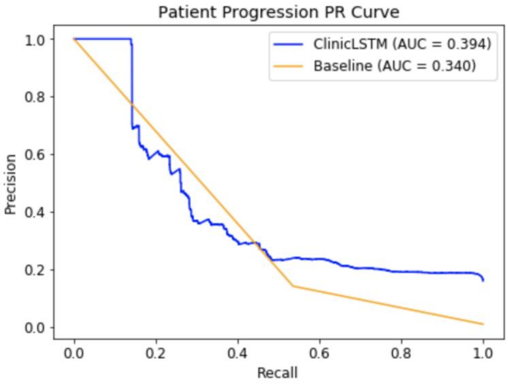

Figure 8: Patient Progression P-R Curve

ClinicNet outperforms existing human-derived order sets as well as other previous models in the literature at predicting what clinical items a physician should order in the next 24 hours. Obtaining high quality order sets can help physicians more efficiently and accurately identify the best clinical items that should be ordered for a patient under their care. In its current state, ClinicNet produces 1,452 clinical item predictions, however, this number could be extended to include any number of desired items.

Further analysis on our subclasses shows that our ClinicNet performs best on recommending nabumetone, amino acids, elastase for stool, acetaminophen-hydrocodone, and albumin with creatinine. However, we perform worse than human authored order sets on consults to nephrology transplants, insulin fasting, testosterone, EC4, and multivitamin preparation-zinc gluconate. One potential reason for this is the smaller set of data that corresponds to human authored order sets that we had (60,000, about 5% of the total dataset), leading to higher bias when training. In addition, it is possible that the top five recommended items are not widely included in today's order sets. Regardless, our model outperforms on over 80% of the clinical items in a scalable and accurate manner.

By transforming the EHR data into patient timelines, we are also able to predict general medical events, or patient progression, using our ClinicLSTM model. ClinicLSTM was able to beat a "Majority" model in precision, F1-score and AUROC, demonstrating the ability of ClinicLSTM to learn from a patient's medical history. However, ClinicLSTM likely suffered from a lack of enough training data and further work on ClinicLSTM will be needed to provide medical professionals with even more accurate and valuable insights into the future path of a given patient through the hospital system.

The success of these models demonstrate the ability of our deep neural networks to understand the patient context. The implications include a scalable and accurate approach to recommending clinical items, as well as predicting patient progression within the hospital. Ultimately, these can help support and guide more effective, evidence-based, and compliant practice.

6 Future Work

As we have demonstrated the efficacy of our clinical decision support model, our future work will include beginning the process for implementing such a system into the clinical flow as well as making further improvements to our system. For instance, while ClinicLSTM outperformed the "Majority" model in precision, F1-score, and AUROC, it did not fare as well on recall. There are a number of adjustments that can be made that will likely improve model performance. First, after generating patient timelines and splitting each timeline into distinct 48-hour pieces, we were left with only 28,859 input sequences. We expect that obtaining more input sequences from EHR will improve the ability of ClinicLSTM to predict patient progression events. Second, adding an attention mechanism to better capture prior a patient's medical history may improve model performance. For our ClinicNet, we hope to implement time-aware attention models, tune hyperparameters for longer periods of time, and collect more data for training.

In addition to the EHR data used in training ClinicNet and ClinicLSTM, there are a wealth of clinical notes written by doctors and nurses pertaining to patient health that may provide insights into future clinical orders and patient progression. Generating embeddings from these notes to be used as input to ClinicNet and ClinicLSTM has the potential to significantly improve the performances of both models.

Another task of interest for our group includes looking at subsets of patients, such as those in the ICU, and seeing if we can provide more accurate, robust, and reliable decision support based on existing medical knowledge when constructing features and response vectors for a more specific task.

7 Conclusion

The current status quo clinical order sets, authored by clinicians, is a largely ineffective, unscalable system for supporting clinical decision making. We developed a deep neural network based model, ClinicNet, that outperformed this system by all metrics. Furthermore, we constructed an LSTM-based recurrent neural network model, called ClinicLSTM, which could predict patient progression outcomes. These models have the potential to provide information that could further inform physicians in their clinical decision-making progress, and save lives. All considered, our work shows the effectiveness of deep learning models in capturing patient context and predicting future events and clinical orders, as well as demonstrates the potential of such models to support the clinical decision making process.

8 Acknowledgements

We would like to acknowledge and thank Dr. Jonathan Chen for advising this project and providing invaluable insights to our team regarding our methodology and approach. Additionally, Santosh Balasubramanian from the Stanford HealthRex lab supported our data pre-processing efforts and Guillaume Genthial (CS230 teaching staff) gave helpful feedback on the LSTM network architecture.

9 Contributions

Jonathan Wang generated the baseline model for the human-authored clinical order sets and compared the results of the human-authored order sets to the two neural network models, as well as querying and obtaining access to data. Delaney Sullivan queried the raw data from the EHR, optimized pre-processing of the raw data, and extracted the feature matrix and response vector from the data. Adam Wells implemented, tuned the hyperparameters, and evaluated the performance of the ClinicNet feed-forward neural network. Alex Wells generated the patient timelines from the raw data and implemented, tuned, and evaluated the ClinicLSTM recurrent neural network. All authors contributed to writing the manuscript.
Code is available for public use here:
https://github.com/HealthRex/CDSS/tree/master
/scripts/DeepLearningRecommender

References

[1] R. Madhok, "Crossing the quality chasm: Lessons from health care quality improvement efforts in england," *Baylor University Medical Center Proceedings*, vol. 15, no. 1, p. 77–83, 2002.

[2] D. T. Durack, "The weight of medical knowledge," *New England Journal of Medicine*, vol. 298, no. 14, pp. 773–775, 1978.

[3] P. Tricoci, "Scientific evidence underlying the acc/aha clinical practice guidelines," *Jama*, vol. 301, no. 8, p. 831, 2009.

[4] S. T. Mennemeyer, N. Menachemi, S. Rahurkar, and E. W. Ford, "Impact of the hitech act on physicians' adoption of electronic health records," *Journal of the American Medical Informatics Association*, vol. 23, no. 2, pp. 375–379, 2016.

[5] J. D. Clough and M. Mcclellan, "Implementing macra," *Jama*, vol. 315, no. 22, p. 2397, 2016.

[6] J. H. Chen, M. K. Goldstein, S. M. Asch, L. Mackey, and R. B. Altman, "Predicting inpatient clinical order patterns with probabilistic topic models vs conventional order sets," *Journal of the American Medical Informatics Association*, 2016.

[7] K. Petersen, M. Nielsen, P. Diao, N. Karssemeijer, and M. Lillholm, "Breast tissue segmentation and mammographic risk scoring using deep learning," *Breast Imaging Lecture Notes in Computer Science*, p. 88–94, 2014.

[8] A. Avati, K. Jung, S. Harman, L. Downing, A. Ng, and N. H. Shah, "Improving palliative care with deep learning," *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2017.

[9] G. Carneiro, J. Nascimento, and A. P. Bradley, “Unregistered multiview mammogram analysis with pretrained deep learning models,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 652–660, Springer, 2015.

[10] Z. Liang, G. Zhang, J. X. Huang, and Q. V. Hu, “Deep learning for healthcare decision making with emrs,” in *Bioinformatics and Biomedicine (BIBM), 2014 IEEE International Conference on*, pp. 556–559, IEEE, 2014.

[11] H. J. Lowe, T. A. Ferris, P. M. Hernandez, and S. C. Weber, “Stride–an integrated standards-based translational research informatics platform,” in *AMIA Annual Symposium Proceedings*, vol. 2009, p. 391, American Medical Informatics Association, 2009.

[12] P. Hernandez, T. Podchiyska, S. Weber, T. Ferris, and H. Lowe, “Automated mapping of pharmacy orders from two electronic health record systems to rxnorm within the stride clinical data warehouse,” in *AMIA Annual Symposium Proceedings*, vol. 2009, p. 244, American Medical Informatics Association, 2009.